AC

(54) Title: IDENTIFICATION AND COMPARISON OF PROTEIN-PROTEIN INTERACTIONS AND INHIBITORS THEREOF

(57) Abstract

Disclosed are methods of detecting protein-protein interactions among two populations of proteins, wherein each protein population has a complexity of at least 1,000. Fusion proteins of each population are expressed in yeast cells of opposite mating types. The fusion protein populations are made by fusing to one population a DNA-binding domain of a transcriptional activator and fusing to the other population at the activation domain of a transcriptional activator. When the yeast cells of opposite mating type are mated, productive interactions between members of each protein population functionally reconstitute the two domains of the transcriptional activator and result in reporter gene expression. The disclosed methods allow identification and characterization of new protein-protein interactions that may be relevant to a particular tissue or disease stage. Inhibitors of the identified protein-protein interactions can also be identified by screening for the ability to reverse expression of reporter gene. This inhibitor screening method can be performed in multiplex. Other aspects of the invention include information processing methods and systems. The methods and systems provide for assembling and processing of a unified database of sequences and identifying sequences that may be involved in protein-protein interactions.

IDENTIFICATION AND COMPARISON OF PROTEIN-PROTEIN INTERACTIONS AND INHIBITORS THEREOF

This application is a continuation in part of co-pending application Serial No. 08/663,824, filed June 14, 1996, which is incorporated by reference herein in its entirety.

This invention was made with United States Government support under award number 70NANB5H1066 awarded by the National Institute of Standards and Technology. The United States Government has certain rights in the invention.

1.   INTRODUCTION

The present method relates to the identification of protein-protein interactions and inhibitors of these interactions that, preferably, are specific to a cell type, tissue type, stage of development, or disease state or stage.

2.   BACKGROUND OF THE INVENTION

Proteins and protein-protein interactions play a central role in the various essential biochemical processes. For example, these interactions are evident in the interaction of hormones with their respective receptors, in the intracellular and extracellular signalling events mediated by proteins, in enzyme substrate interactions, in intracellular protein trafficking, in the formation of complex structures like ribosomes, viral coat proteins, and filaments, and in antigen-antibody interactions. These interactions are usually facilitated by the interaction of small regions within the proteins that can fold independently of the rest of the protein. These independent units are called protein domains. Abnormal or disease states can be the direct result of aberrant protein-protein interactions. For example, oncoproteins can cause cancer by interacting with and activating proteins responsible for cell division. Protein-protein interactions are also central to the mechanism of a virus recognizing its receptor on the cell

- 1 -

surface as a prelude to infection. Identification of domains that interact with each other not only leads to a broader understanding of protein-protein interactions, but also aids in the design of inhibitors of these interactions.

5   Protein-protein interactions have been studied by both biochemical and genetic methods. The biochemical methods are laborious and slow, often involving painstaking isolation, purification, sequencing and further biochemical characterization of the proteins being tested for

10 interaction. As an alternative to the biochemical approaches, genetic approaches to detect protein-protein interactions have gained in popularity as these methods allow the rapid detection of the domains involved in protein-protein interactions.

15   An example of a genetic system to detect protein-protein interactions is the "Two-Hybrid" system to detect protein-protein interactions in the yeast Saccharomyces cerevisiae (Fields and Song, 1989, Nature 340:245-246; U.S. Patent No. 5,283,173 by Fields and Song). This assay

20 utilizes the reconstitution of a transcriptional activator like GAL4 (Johnston, 1987, Microbiol. Rev. 51:458-476) through the interaction of two protein domains that have been fused to the two functional units of the transcriptional activator: the DNA-binding domain and the activation domain.

25 This is possible due to the bipartite nature of certain transcription factors like GAL4. Being characterized as bipartite signifies that the DNA-binding and activation functions reside in separate domains and can function in trans (Keegan et al., 1986, Science 231:699-704). The

30 reconstitution of the transcriptional activator is monitored by the activation of a reporter gene like the *lacZ* gene that is under the influence of a promoter that contains a binding site (Upstream Activating Sequence or UAS) for the DNA-binding domain of the transcriptional activator. This method

35 is most commonly used either to detect an interaction between two known proteins (Fields and Song, 1989, Nature 340:245-246) or to identify interacting proteins from a

- 2 -

population that would bind to a known protein (Durfee et al.,
1993, Genes Dev. 7:555-569; Gyuris et al., 1993, Cell
75:791-803; Harper et al., 1993, Cell 75:805-816; Vojtek et
al., 1993, Cell 74:205-214).

5   Another system that is similar to the Two-Hybrid
system is the "Interaction-Trap system" devised by Brent and
colleagues (Gyuris et al., 1993, Cell 75:791-803). This
system is similar to the Two-Hybrid system except that it
uses a *LEU2* reporter gene and a *lacZ* reporter gene. Thus
10 protein-protein interactions leading to the reconstitution of
the transcriptional activator also allow cells to grow in
media lacking leucine and enable them to express
β-galactosidase. The DNA-binding domain used in this system
is the LexA DNA-binding domain, while the activator sequence
15 is obtained from the B42 transcriptional activation domain
(Ma and Ptashne, 1987, Cell 51:113-119). The promoters of
the reporter genes contain LexA binding sequences and hence
will be activated by the reconstitution of the
transcriptional activator. Another feature of this system is
20 that the gene encoding the DNA-binding domain fusion protein
is under the influence of an inducible GAL promoter so that
confirmatory tests can be performed under inducing and non-
inducing conditions.

           In yet another version of this system developed by
25 Elledge and colleagues, the reporter genes *HIS3* and *lacZ*
(Durfee et al., 1993, Genes Dev. 7:555-569) are used. The
transcriptional activator that is reconstituted in this case
is GAL4 and protein-protein interactions allow cells to grow
in media lacking histidine and containing 3-aminotriazole
30 (3-AT) and to express β-galactosidase. 3-AT inhibits the
growth of *his3* auxotrophs in media lacking histidine (Kishore
and Shah, 1988, Annu. Rev. Biochem. 57:627-663).

           In a different two-hybrid assay, a *URA3* reporter
gene under the control of Estrogen Response Elements (ERE)
35 has been used to monitor protein-protein interactions. Here,
the DNA-binding domain is derived from the human estrogen
receptor. The authors of the ERE assay propose that

- 3 -

inhibition of the protein-protein interactions can be
identified by negative selection on 5-FOA medium (Le Douarin
et al., 1995, Nucleic Acids Res. 23:876-878), but do not
provide any details.

5   A version of the two-hybrid approach called the
"Contingent Replication Assay" that is applicable in
mammalian cells has also been reported (Nallur et al., 1993,
Nucleic Acids Res. 21:3867-3873; Vasavada et al., 1991, Proc.
Natl. Acad. Sci. USA 88:10686-10690). In this case, the
10 reconstitution of the transcription factor in mammalian cells
due to the interaction of the two fusion proteins leads to
the activation of the SV40 T antigen. This antigen allows
the replication of the activation domain fusion plasmids.
Another modification of the two-hybrid approach using
15 mammalian cells is the "Karyoplasmic Interaction Selection
Strategy" that also uses the reconstitution of a
transcriptional activator (Fearon et al., 1992, Proc. Natl.
Acad. Sci. USA 89:7958-7962). Reporter genes used in this
case have included the gene encoding the bacterial
20 chloramphenicol acetyl transferase, the gene for cell-surface
antigen CD4, and the gene encoding resistance to Hygromycin
B. In both of the mammalian systems, the transcription
factor that is reconstituted is a hybrid transcriptional
activator in which the DNA-binding domain is from GAL4 and
25 the activation domain is from VP16.

          In all of the assays described above, the identity
of one (or both) of the proteins being tested for interaction
is known. All of the assays mentioned above can be used to
identify novel proteins that interact with a known protein of
30 interest. In a variation of the "Interaction Trap" system, a
"mating-grid" strategy has been used to characterize
interactions between proteins that are thought to be involved
in the Drosophila cell cycle (Finley and Brent, 1994, Proc.
Natl. Acad. Sci. USA 91:12980-12984). This strategy is based
35 on a technique first established by Rothstein and colleagues
(Bendixen et al., 1994, Nucleic Acids Res. 22:1778-1779) who
used a yeast-mating assay to detect protein-protein

- 4 -

interactions. Here, the DNA-binding and activation domain
fusion proteins were expressed in two different haploid yeast
strains, a and α, and the two were brought together by
mating. Thus, interactions between proteins can be studied
5 in this method. However, even in this method, the identities
of at least one of the proteins in the interacting pairs of
proteins was known prior to analyzing the interactions
between pairs of proteins.

Stanley Fields and coworkers have recently
10 performed an analysis of all possible protein-protein
interactions that can take place in the *E. coli* bacteriophage
T7 (Bartel et al., 1996, Nature Genet. 12:72-77). Randomly
sheared fragments of T7 DNA were used to make libraries in
both the DNA-binding domain and the activation domain
15 plasmids and a genome-wide two-hybrid assay was performed by
use of a mating strategy. The DNA-binding and the activation
domain fusions were transformed into separate yeast strains
of opposite mating type. The DNA-binding domain hybrids
containing yeast transformants were then divided into groups
20 of 10. The groups were screened (by the mating strategy
outlined above) against a library of activation domain
hybrids numbering around 10^5 transformants. By this method,
25 interactions were characterized among the proteins of T7.
While this study provides a method to screen more than one
25 DNA-binding domain hybrid against more than one activation
domain hybrid, it does not address the issues involved in
screening complex libraries against each other. This is an
important limitation due to the value of enabling the
detection and isolation of interactants from cDNA libraries
30 prepared from complex organisms like human beings. Indeed,
the prior art has taught away from using complex populations
of proteins as hybrids to the DNA-binding domain, since
random hybrids to the DNA binding domain produce a large
percentage of false positives (hybrids that have
35 transcriptional activity in the absence of an interacting
protein) (Bartel et al., 1993, "Using the two hybrid system
to detect protein-protein interactions," in *Cellular*

- 5 -

Transduction in Development, Ch. 7, Hartley, D.A. (ed.),
Practical Approach Series xviii, IRL Press at Oxford
University Press, New York, NY, pp. 154-179 at 171; Ma and
Ptashne, 1987, Cell 51:113).

5   None of the prior art systems provides a method
that not only isolates and catalogues all possible protein-
protein interactions within a population, be it a
tissue/cell-type, disease state, or stage of development, but
also allows the comparison of such interactions between two
10 such populations thereby allowing the identification of
protein-protein interactions unique to any particular
tissue/cell-type, disease state, or stage of development.   In
contrast, such a method is provided by the present invention.
           Accordingly, it is one of the objectives of this
15 invention to devise a genetic method to identify and isolate
preferably all possible protein-protein interactions within a
population of proteins, or between two different populations
of proteins, be it a tissue/cell-type, disease state or stage
of development.

20   It is another objective of the present invention to
perform a comparative analysis of the protein-protein
interactions that occur two or more different tissue/cell-
types, disease states, or stages of development.
           It is also an objective of this invention to
25 identify and isolate in a rapid manner the genes encoding the
proteins involved in interactions that are specific to a
tissue/cell-type, disease state, or stage of development.
           It is yet another objective of this invention to
provide a method for the concurrent identification of
30 inhibitors of the protein-protein interactions that
characterize a given population, be it a tissue/cell type,
disease state, or stage of development.   These inhibitors may
have therapeutic value.
           Citation of a reference herein shall not be
35 construed as an admission that such is prior art to the
present invention.

3. SUMMARY OF THE INVENTION

The present invention provides methods and means to detect and isolate the genes encoding the proteins that interact with each other between two populations of proteins,

5 using the reconstitution of a selectable event. This selectable event is the formation of a transcription factor. In contrast to the prior art, in which problems with false positives and low throughput limited the complexity of the populations that could be analyzed, each of the two

10 populations of proteins has a complexity of greater than 10, and preferably has a complexity of at least 1,000. The reconstitution of a transcription factor occurs by interaction of fusion proteins expressed by chimeric genes. In a preferred embodiment, the types of fusion proteins used

15 are DNA-binding domain hybrids and activation domain hybrids of transcriptional activators. Libraries of genes encoding hybrid proteins are preferably constructed in both a DNA-binding domain hybrid plasmid vector and in an activation domain hybrid plasmid vector. In a preferred embodiment, two

20 types of haploid yeast strains, a and α respectively, are each transformed with a different one of the two libraries to create two yeast libraries. The two yeast libraries are then mated together to create a diploid yeast strain that contains both the kinds of fusion genes encoding the hybrid proteins.

25 If the two hybrid proteins can interact (bind) with each other, the transcriptional activator is reconstituted due to the proximity of the DNA-binding and the activation domains of the transcriptional activator. This reconstitution causes transcription of reporter genes that, by way of example,

30 enable the yeast to grow in selective media. In a preferred aspect, the activity of a reporter gene is monitored enzymatically. The isolation of the plasmids that encode these fusion genes leads to the identification of the genes that encode proteins that interact with each other.

35   Thus, in a specific embodiment, the invention is directed to a method of detecting one or more protein-protein interactions comprising (a) recombinantly expressing within

a population of host cells (i) a first population of first
fusion proteins, each said first fusion protein comprising a
first protein sequence and a DNA binding domain in which the
DNA binding domain is the same in each said first fusion
5 protein, and in which said first population of first fusion
proteins has a complexity of at least 1,000; and (ii) a
second population of second fusion proteins, each said second
fusion protein comprising a second protein sequence and a
transcriptional regulatory domain of a transcriptional
10 regulator, in which the transcriptional regulatory domain is
the same in each said second fusion protein, such that a
first fusion protein is co-expressed with a second fusion
protein in host cells, and wherein said host cells contain at
least one nucleotide sequence operably linked to a promoter
15 driven by one or more DNA binding sites recognized by said
DNA binding domain such that interaction of a first fusion
protein with a second fusion protein results in regulation of
transcription of said at least one nucleotide sequence by
said regulatory domain, and in which said second population
20 of second fusion proteins has a complexity of at least 1,000;
and (b) detecting said regulation of transcription of said at
least one nucleotide sequence, thereby detecting an
interaction between a first fusion protein and a second
fusion protein.
25   In further specific embodiments, this invention
provides for detecting experimentally significant protein-
protein interactions between highly complex libraries of
proteins. In particular, the invention provides protocols
which achieve highly effective screening of the DNA binding
30 domain or activation domain hybrids to eliminate those
hybrids that produce false positive indications of protein-
protein interactions. Additional screening protocols
eliminate those hybrids which, due to non-specific
association with many proteins, produce less experimentally
35 significant or specific indications of protein-protein
interactions. Further protocols provide for the efficient

mating of large numbers of yeast cells useful for handling complex libraries.

The present invention also provides a method to isolate concurrently inhibitors of such protein-protein
5   interactions that occur in, are characteristic of or are specific to a given population of proteins. By way of example, preferably all the yeast diploids that harbor fusion proteins that interact with each other are pooled together and exposed to candidate inhibitors. Exemplary candidate
10 inhibitors include chemically synthesized molecules and genetically encoded peptides. After treatment with candidate inhibitors, the yeast cells harboring interacting hybrid proteins are selected for the inactivation of the reporter gene, preferably by transfer to appropriate selective media.
15 Preferably, the same media also selects for the presence of the plasmids that encode the interacting proteins, and the peptide-encoding peptides in the case of the screening for peptide inhibitors expressed from expression plasmids. Successful inhibition events are thus monitored by the
20 inactivation of the reporter gene.

The major advantages of these methods are as follows. From a population of proteins characteristic of a particular tissue or cell-type, all possible detectable protein-protein interactions that occur can be identified and
25 the genes encoding these proteins can be isolated. Thus, parallel analyses of two cell types enumerates the protein-protein interactions that are common to both and those that are specific to both (differentially expressed in one cell type and not the other). Such an analysis has value since
30 protein-protein interactions specific to a disease state can serve as therapeutic points of intervention.

Furthermore, inhibitors of such protein-protein interactions can be isolated in a rapid fashion. Such inhibitors can be of therapeutic value or serve as lead
35 compounds for the synthesis of therapeutic compounds. This system can also be used to identify novel peptide inhibitors of protein-protein interactions. One advantage of this

method over existing methods is that peptides or chemicals
are identified by an ability to block protein-protein
interactions. In many existing methods, molecules are
identified by an ability only to bind to one of a pair of
5 interacting proteins; such binding does not necessarily imply
that the protein-protein interaction will be blocked by the
same agent. Another advantage of the method is that multiple
protein-protein interactions can be screened against a
prospective inhibitor in a single assay.
10   This invention also provides information-processing
methods and systems. One aspect of these methods provides
methods for interpreting detected protein-protein
interactions by providing for identification of the genes
that code for the library inserts in the activation domain
15 and fusion domain hybrids. Another aspect of these methods
provides for assembling protein-protein interaction data
detected from one or more pairs of libraries into a unified
database. Further aspects of these methods provide for use
of this unified database to assemble individual, pair-wise
20 protein-protein interactions into putative pathways and
networks of protein interaction, providing a more general
view of cellular functioning. Also provided for is the use
of this unified database to delimit or determine the protein
domains responsible for particular protein-protein
25 interactions.

4. DESCRIPTION OF THE FIGURES

These and other features, aspects, and advantages
of the present invention will become better understood by
30 reference to the accompanying drawings, following
description, and appended claims, where the drawings are
described briefly as follows:
Figure 1. An overview of an exemplary strategy to
identify pairs of interacting proteins that are specific to a
35 particular population and to identify inhibitors of these
interactors in a high throughput fashion.

Figure 2.   A yeast interaction mating assay for
the detection of protein-protein interactions. The two test
proteins are indicated as X and Y. X = DNA binding domain
fusion protein; Y = activation domain fusion protein. The
5 activation and DNA-binding domains are indicated as A and D
respectively. The two yeast cell types are a and α, while
the diploid is marked as a/α. A blue color (not shown)
indicates expression of β-galactosidase by conversion of the
clear X-gal substrate into an insoluble blue precipitate.

10   Figure 3. Exemplary scheme for the isolation of
stage-specific pairs of interacting proteins. M and N are
two populations of proteins expressed in a particular state
(e.g., cancer). The mating of two populations M and N
results in the creation of an interactive population that
15 contains all possible pairs of interacting proteins in the
two populations. The reporter genes are URA3, HIS3, and
lacZ. The interactive population is further characterized by
methods such as the QEA™ method, the SEQ-QEA™ method, and
sequencing which aid in the identification of the pairs of
20 interacting proteins. A comparison of two such interactive
populations leads to the identification of stage-specific or
disease state-specific pairs of interacting proteins.

          Figures 4A-C. Pooling strategies to characterize
the interactive populations. PCR reactions are performed on
25 pooled yeast cells and the PCR products are either analyzed
directly by electrophoresis or by the QEA™ method and SEQ-
QEA™ method. These methods lead to the characterization of
an interactive population. (A) 2-dimensional pooling and
deconvolution. (B) 3-dimensional pooling and deconvolution.
30 In order to determine the location of each clone, wells are
pooled along planes (as opposed to lines in the 2-dimensional
strategy). The location of a specific gene can be determined
by finding which pool from each axis contains it. (C)
3-dimensional pooling from 96 well plates. 1152 positive
35 colonies are arrayed into individual wells of twelve
microtiter plates. A total of 32 pools are produced: 12
pooled along the columns axis (each from all 12 plates), 8

- 11 -

pooled along rows (A-H), and 12 pooled plates (p1-p12).
These pools contain genes from 96, 144, and 96 wells,
respectively. Two-dimensional pooling and deconvolution
requires 36 + 24 pools, but no pool is from more than 36
5 wells (genes), so it is easier to get clearly separate bands
from a SEQ-QEA™ method reaction of pools than with the three-
dimensional strategy.

Figure 5. Isolation of stage-specific pairs of
interacting proteins by probing interactive grids. M and N
10 are two populations of proteins expressed in a particular
state (e.g., cancer). The PCR products corresponding to M
and N partners from an M x N analysis are spotted onto a
solid support like a nitrocellulose membrane to create an
interactive grid. The interactive grids are then probed with
15 DNA that is unique to a specific stage to isolate pairs of
interacting proteins that are unique to a specific stage.

Figure 6. Integration of the expression linkage
analysis and inhibitor screen. Exemplary steps in an
integrated isolation of inhibitors of protein-protein
20 interactions are depicted. The interactive populations that
arise from an M x N analysis are screened against many
inhibitors such that only successful inhibition events are
selected. Thus, from an M x N analysis not only are obtained
stage-specific pairs of interacting proteins, but also
25 inhibitors of such interactions.

Figure 7. Peptide expression vector polylinker.
The polylinker region of the peptide expression vector (PEV)
is depicted. ADC1-P and ADC1-T refer to the ADC1 promoter
and terminator, respectively. This is a yeast promoter that
30 promotes transcription of genes downstream of it. Sfi I and
Asc I sites demarcate the region within which the peptide-
coding regions will be inserted. UAG refers to the
termination codon and NLS refers to the Nuclear Localization
Signal that provides transport of the peptides into the
35 nucleus.

Figure 8. A QEA™ Method Analysis. A comparison is
depicted of a QEA™ method pattern from an M x N analysis

conducted in duplicate (Section 6.5). The PCR products that were pAS-like vector-specific were pooled and subjected to a QEA™ method analysis. I and II refer to duplicate M x N analyses. The dotted peaks correspond to the molecular
5 weight markers and the solid peaks are the QEA™ method products.

Figure 9. The QEA™ Method Comparison. A comparison is shown of the QEA™ method patterns from an M x N analysis conducted wherein one of the interactive populations
10 had the RAS-RAF interaction. The RAF peak obtained in the QEA™ method is shown in solid black.

Figures 10A, 10B, 10C, and 10D show DNA adapters for an RE/ligation implementation of a Quantitative Expression Analysis ("QEA™") method, where the restriction
15 endonucleases generate 5' overhangs, open blocks indicating strands of DNA;

Figures 11A and 11B show the DNA adapters for an RE/ligation implementation of a QEA™ method, where the restriction endonucleases generate 3' overhangs;
20 Figures 12A, 12B, and 12C show an exemplary biotin alternative embodiment of the QEA™ method;

Figures 13A and 13B show a method for DNA sequence database selection according to a QEA™ method;

Figure 14 shows an exemplary experimental
25 description for an embodiment of a QEA™ method;

Figures 15A and 15B show an overview of a method for determining a simulated database of experimental results for an embodiment of a QEA™ method;

Figure 16 shows the detail of a method for
30 simulating a QEA™ reaction;

Figure 17A-F show exemplary results of the action of the method of Figure 16;

Figure 18 shows the detail of a method for determining a simulated database of experimental results for
35 a QEA™ embodiment;

Figures 19A and 19B show an exemplary computer system apparatus implementing the QEA™ methods;

- 13 -

Figure 20A shows exemplary detail of an experimental design method for a QEA™ method, and Figure 20B shows exemplary detail of an experimental design method for an embodiment of a QEA™ method;

5   Figure 21 shows an exemplary method for ordering the DNA sequences found to be likely causes of a QEA™ method signal in the order of their likely presence in the sample;

Figures 22A, 22B, 22C, and 22D show exemplary reaction temperature profiles for preferred manual and

10 automated implementations of a preferred RE embodiment of a QEA™ method; and

Figures 23A-23E show details of a SEQ-QEA™ embodiment of a QEA™ method.

Figure 24.   Exemplary protocol for selection of

15 inhibitors of protein-protein interactions.

Figure 25.   Exemplary protocol for selection of novel interacting proteins and inhibitors of these interacting proteins.

Figure 26.   Exemplary method steps for a particular

20 alternative embodiment for detecting protein-protein interactions and exemplary information processing steps.

Figure 27.   Exemplary computer-implemented system for performing the information processing steps of Figure 26.

Figures 28A and 28B.   Exemplary computer display

25 screens for data selection according to the information processing steps of Figure 26.

Figure 29.   Exemplary computer display screen for protein interaction pathways according to the information processing steps of Figure 26.

30   Figure 30.   Example of an exemplary method for finding domains responsible for interaction according to the information processing steps of Figure 26.

5.   <u>DETAILED DESCRIPTION OF THE INVENTION</u>

35   In contrast to prior art methods of detecting protein-protein interactions between two protein populations, wherein the number of false positives and low throughput

- 14 -

limited the applicability of such prior art methods to
situations in which the complexity of at least one of the
populations was no more than 10, the present invention allows
detection of protein-protein interactions (and isolation and
5 characterization of the interacting proteins) between
populations in which both populations can have complexities
of orders of magnitude significantly greater than 10, e.g.,
1,000, 100,000, or in the range of 50,000-100,000 as is found
in mammalian cDNA populations. Methods for detecting,
10 isolating, and characterizing inhibitors of such interactions
are also provided.

   For purposes of convenience of description and not
by way of limitation, the detailed description is divided
into the subsections set forth below.
15

5.1. DETECTING INTERACTING PROTEINS

   The present invention provides methods for
detecting interacting proteins (including peptides).
Interacting proteins are detected based on the reconstitution
20 of a transcriptional regulator in the presence of a reporter
gene ("Reporter Gene") whose transcription is then regulated
by the reconstituted regulator. In contrast to prior art
methods, the protein-protein interactions can be detected,
and the interacting pairs of proteins isolated and
25 identified, between two populations of proteins wherein both
of the populations have a complexity of at least 10 (i.e.,
both populations contain more than ten distinct proteins).
The populations are expressed as fusion proteins to a DNA
binding domain, and to a transcriptional regulatory domain,
30 respectively. In various specific embodiments, one or both
of the populations of proteins has a complexity of at least
50, 100, 500, 1,000, 5,000, 10,000, or 50,000; or has a
complexity in the range of 25 to 100,000, 100 to 100,000,
50,000 to 100,000, or 10,000 to 500,000. For example, one or
35 both populations can be mammalian cDNA populations, generally
having a complexity in the range of 50,000 to 100,000; in
such populations from total mRNA, the detection of a protein

- 15 -

in an interacting pair that is expressed to a particular
level can be optimized by the statistical considerations
described in Section 5.2.7 below. In a specific embodiment,
the invention is capable of detecting substantially all
5 detectable interactions that occur between the component
proteins of two populations, each population having a
complexity of at least 50, 100, 500, 1000, 5000, 10,000 or
50,000. In a specific embodiment, the two populations are
samples (aliquots) of at least 100 or 1000 members (e.g.,
10 expressed in host yeast cells) of a larger population (e.g.,
a mammalian cDNA library) having a complexity of at least
100, 1000, 5,000, 10,000, or 50,000; in a particular
embodiment, the sample is uncharacterized in that the
particular identities of all or most of its member proteins
15 are not known.

The populations can be the same or different
populations. If it is desired to detect interactions between
proteins encoded by a particular DNA population, both protein
populations are expressed from chimeric genes comprising DNA
20 sequences representative of that particular DNA population.
In another embodiment, one protein population is expressed
from chimeric genes comprising cDNA sequences of diseased
human tissue, and the other protein population is expressed
from chimeric genes comprising cDNA sequences of non-diseased
25 human tissue. In a specific embodiment, one or more of the
populations can be uncharacterized in that the identities of
all or most of the members of the population are not known.
Preferably, the populations are proteins encoded by DNA,
e.g., cDNA or genomic DNA or synthetically generated DNA.
30 For example, the populations can be expressed from chimeric
genes comprising cDNA sequences from an uncharacterized
sample of a population of cDNA from mammalian RNA.
Preferably, a cDNA library is used. The cDNA can be, e.g., a
normalized or subtracted cDNA population. The cDNA of one or
35 both populations can be cDNA of total mRNA or polyA⁺ RNA or a
subset thereof from a particular species, particular cell
type, particular age of individual, particular tissue type,

- 16 -

disease state or disorder or stage thereof, or stage of
development. Accordingly, the invention provides methods of
identifying and isolating interacting proteins that are
present in or specific to particular species, cell type, age,
5 tissue type, disease state, or disease stage, and also
provides methods for comparing the protein-protein
interactions present in such particular species, cell type,
age, tissue type, disease state, or disease stage (by e.g.,
using a cDNA library of total mRNA particular to such
10 species, cell type, age, tissue type, disease state, or
disease stage, respectively, as both the populations between
which interactions are detected) with the protein-protein
interactions present in a different species, cell type, age,
tissue type, non-diseased state or a different disease stage,
15 or different state of development, respectively. For
example, in one embodiment, interactions are detected between
identical populations of proteins in which the population of
proteins is from cDNA of cancerous or precancerous (e.g.,
hyperplastic, metaplastic, or dysplastic cells), e.g., of
20 prostate cancer, breast cancer, stomach cancer, lung cancer,
ovarian cancer, uterine cancer, etc.; these interactants are
then compared to interacting proteins detected between two
other identical populations of proteins in which the
population of proteins is from cDNA of cells not having the
25 cancer or precancerous condition, as the case may be. In a
specific embodiment, cDNA may be obtained from a preexisting
cDNA sample or may be prepared from a tissue sample. When
cDNA is prepared from tissue samples, methods commonly known
in the art can be used. For example, these can consist of
30 largely conventional steps of RNA preparation from the tissue
sample, preferably total poly(A) purified RNA is used but
less preferably total cellular RNA can be used, RNase
extraction, DNase treatment, mRNA purification, and first and
second strand cDNA synthesis.
35   Preferably, the populations of proteins between
which interactions are detected are provided by recombinant
expression of nucleic acid populations (e.g., cDNA or genomic

- 17 -

libraries). Also preferably, the interactions occur
intracellularly. In another specific embodiment, recombinant
biological libraries expressing random peptides can be used
as the source nucleic acid for one or both of the nucleic
5 acid populations.

In a specific embodiment, presented by way of
example and not limitation, the method of the invention
comprises the steps schematically depicted in Figure 1.

In a preferred aspect, the present invention
10 provides a method for detecting unique protein-protein
interactions that characterize a population or library of
proteins by comparing all detectable protein-protein
interactions that occur in a population or library with those
interactions that occur in another population or library.
15 Furthermore, the method also enables the identification of
inhibitors of such protein-protein interactions.

Protein-protein interactions are detected according
to the invention by detecting transcriptional regulation
(preferably activation) which occurs upon interaction of
20 proteins between the two populations being tested (referred
to hereinafter merely for purposes of convenience as the M
population and the N population). Proteins of each
population (M, N) are provided as fusion (chimeric) proteins
(preferably by recombinant expression of a chimeric coding
25 sequence) containing each protein contiguous to a preselected
sequence. For one population, the preselected sequence is a
DNA binding domain. The DNA binding domain can be any
available, as long as it specifically recognizes a DNA
sequence within a promoter. For example, the DNA binding
30 domain is of a transcriptional activator or inhibitor. For
the other population, the preselected sequence is an
activator or inhibitor domain of a transcriptional activator
or inhibitor, respectively.

In a preferred embodiment, each protein in one
35 population (e.g., M) is provided as a fusion to a DNA binding
domain of a transcriptional regulator (e.g., activator).
Each protein in the other population (N) is provided as a

- 18 -

fusion to an activator domain of a transcriptional activator. The regulatory domain alone (not as a fusion to a protein sequence) and the DNA-binding domain alone (not as a fusion to a protein sequence) preferably do not detectably interact
5 (so as to avoid false positives in the assay). When binding occurs of a fusion protein in M to a fusion protein in N, reconstitution of a transcriptional activator occurs such that transcription is increased of a gene ("Reporter Gene") responsive to (whose transcription is under the control of)
10 the transcriptional activator. Thus, the Reporter Gene comprises a nucleotide sequence operably linked to a promoter regulated by a DNA binding site for the DNA binding domain of the transcriptional activator. The activation of transcription of the Reporter Gene occurs intracellularly,
15 e.g., in prokaryotic or eukaryotic cells, preferably in cell culture.

The Reporter Gene comprises a nucleotide sequence operably linked to a promoter that is operably linked to one or more nucleic acid binding sites that are specifically
20 bound by the DNA binding domain of the fusion protein that is employed in the assay of the invention, such that binding of a reconstituted transcriptional activator or inhibitor to the one or more DNA binding sites increases or inhibits, respectively, transcription of the nucleotide sequence under
25 the control of the promoter. The promoter that is operably linked to the nucleotide sequence can be a native or non-native promoter of the nucleotide sequence, and the DNA binding site(s) that are recognized by the DNA binding domain portion of the fusion protein can be native to the promoter
30 (if the promoter normally contains such binding site(s)) or non-native. Thus, for example, one or more tandem copies (e.g., 4 or 5 copies) of the appropriate DNA binding site can be introduced upstream of the TATA box in the desired promoter (e.g., in the area of position -100 to -400). In a
35 preferred aspect, 4 or 5 tandem copies of the 17 bp UAS (GAL4 DNA binding site) are introduced upstream of the TATA box in the desired promoter, that is in turn upstream of the desired

coding sequence that encodes a selectable or detectable
marker.   In a preferred embodiment, the *GAL1-10* promoter is
operably fused to the desired nucleotide sequence; the
GAL1-10 promoter already contains 5 binding sites for GAL4.

5 Thus, in a particular embodiment, the transcriptional
activation binding site of the desired gene(s) can be deleted
and replaced with GAL4 binding sites (Bartel et al., 1993,
BioTechniques 14(6):920-924; Chasman et al., 1989, Mol. Cell.
Biol. 9:4746-4749).   Referring to use of a particular gene as

10 a Reporter Gene herein thus means that, if the native
promoter is not driven by binding site(s) recognized by the
DNA binding domain used in the interaction assay of the
invention, such DNA binding site(s) have been introduced into
the gene.

15   The Reporter Gene preferably comprises a nucleotide
sequence, whose transcription is regulated by the
transcriptional activator, that is a coding sequence that
encodes a detectable marker or selectable marker,
facilitating detection of transcriptional activation, thereby

20 detecting a protein-protein interaction.   Preferably, the
assay is carried out in the absence of background levels of
the transcriptional activator (*e.g.*, in a cell that is mutant
or otherwise lacking in the transcriptional activator).
Preferably, more than one different Reporter Gene is used to

25 detect transcriptional activation, *e.g.*, one encoding a
detectable marker, and one or more encoding different
selectable markers.   The detectable marker can be any
molecule that can give rise to a detectable signal, e.g., an
enzyme or fluorescent protein.   The selectable marker can be

30 any molecule which can be selected for its expression, *e.g.*,
which gives cells a selective advantage over cells not having
the selectable marker under appropriate (selective)
conditions.   In preferred aspects, the selectable marker is
an essential nutrient in which the cell in which the

35 interaction assay occurs is mutant or otherwise lacks or is
deficient, and the selection medium lacks such nutrient.   The
Reporter Gene used need not be a gene containing a coding

sequence whose native promoter contains a binding site for
the DNA binding protein, but can alternatively be a chimeric
gene containing a sequence that is transcribed under the
control of a promoter that is not the native promoter for the
5 transcribed sequence.

        In a specific embodiment, to make the fusion
constructs (encoding the fusion proteins such that the fusion
proteins are expressed in the desired host cell) from each
population (e.g., library), the activation domain and DNA
10 binding domain of a wide variety of transcriptional activator
proteins can be used, as long as these transcriptional
activators have separable binding and transcriptional
activation domains. For example, the GAL4 protein of *S.
cerevisiae*, the GCN4 protein of *S. cerevisiae* (Hope and
15 Struhl, 1986, Cell 46:885-894); the ARD1 protein of *S.
cerevisiae* (Thukral et al., 1989, Mol. Cell. Biol.
9:2360-2369), and the human estrogen receptor (Kumar et al.,
1987, Cell 51:941-951) have separable DNA binding and
activation domains. The DNA binding domain and activation
20 domain that are employed in the fusion proteins need not be
from the same transcriptional activator. In a specific
embodiment, a GAL4 or LEXA DNA binding domain is employed.
In another specific embodiment, a GAL4 or herpes simplex
virus VP16 (Triezenberg et al., 1988, Genes Dev. 2:730-742)
25 activation domain is employed. In a specific embodiment,
amino acids 1-147 of GAL4 (Ma et al., 1987, Cell 48:847-853;
Ptashne et al., 1990, Nature 346:329-331) is the DNA binding
domain, and amino acids 411-455 of VP16 (Triezenberg et al.,
1988, Genes Dev. 2:730-742; Cress et al., 1991, Science
30 251:87-90) is the activation domain.

        In a preferred embodiment, the transcriptional
activator that is reconstituted in the manner described above
is the yeast transcription factor GAL4 (Figure 2). The host
strain bears a mutant GAL4 gene (e.g., having a deletion or
35 point mutation) and as such cannot express the GAL4
transcriptional activator.

- 21 -

In another embodiment, the DNA-binding domain is
Ace1N, the DNA binding domain of the Ace1 protein. In
another embodiment, the activation domain is Ace1C, the
activation domain of Ace1. Ace1 is a yeast protein that
5 activates transcription from the CUP1 operon in the presence
of divalent copper. CUP1 encodes metallothionein, which
chelates copper; thus, CUP1 gene expression is Reporter Gene
expression suitable for use with Ace1N, in which selection is
carried out by using copper in the media of the growing host
10 cells which would otherwise be toxic to the cells.
Alternatively or additionally, the Reporter Gene can comprise
a CUP1-lacZ fusion such that the enzyme β-galactosidase is
expressed upon binding of a transcriptional activator
reconstituted with Ace1N (see Chaudhuri et al., 1995, FEBS
15 Letters 357:221-226).

In another specific embodiment, the DNA binding
domain of the human estrogen receptor is used, with a
Reporter Gene driven by one or three estrogen receptor
response elements (see Le Douarin et al., 1995, Nucl. Acids.
20 Res. 23:876-878).

In an embodiment in which the interaction assay is
carried out in a prokaryotic cell and in which fusion
proteins to a transcriptional inhibition domain are used as
one of the populations of proteins, both the DNA binding
25 domain fusion population and the inhibition domain fusion
population can be fusions to the λ cI repressor. In this
embodiment, interaction of two fusion proteins via the non-cI
protein portions promotes oligomerization of the λ cI DNA
binding domain sufficient to cause DNA binding and inhibition
30 of transcription from the two phage major early promoters,
preventing lytic growth and rendering the host bacterial
cells immune to superinfection by λ (Hu et al., 1995,
Structure 3:431-433). Alternatively, the DNA binding domains
of the LexA repressor (Schmidt-Dörr et al., 1991,
35 Biochemistry 30:9657-9664), 434 repressor (Pu et al., 1993,
Nucl. Acids Res. 21:4348-4355), or AraC protein (Bustos et
al., 1993, Proc. Natl. Acad. Sci. USA 90:5638-5642) can be

used in both the DNA binding domain and the transcription
inhibition fusion populations.

　　　　The DNA binding domain and the transcription
activator/inhibitor domain each preferably comprises a
5 nuclear localization signal (see Ylikomi et al., 1992, EMBO
J. 11:3681-3694; Dingwall and Laskey, 1991, TIBS 16:479-481)
functional in the cell in which the fusion proteins are to be
expressed.

　　　　In another embodiment, the fusion constructs
10 further comprise sequences encoding affinity tags such as
glutathione-S-transferase or maltose-binding protein or an
epitope of an available antibody, so as to facilitate
isolation of the encoded proteins by affinity methods (e.g.,
binding to glutathione, maltose, or antibody, respectively)
15 (see Allen et al., 1995, TIBS 20:511-516). In another
embodiment, the fusion constructs further comprise bacterial
promoter sequences operably linked to the fusion coding
sequences to facilitate the production of the fusion proteins
also in bacterial cells (see Allen et al., 1995, TIBS
20 20:511-516).

　　　　The host cell in which the interaction assay occurs
can be any cell, prokaryotic or eukaryotic, in which
transcription of the Reporter Gene can occur and be detected,
including but not limited to mammalian (e.g., monkey,
25 chicken, mouse, rat, human, bovine), bacteria, and insect
cells, and is preferably a yeast cell. Expression constructs
encoding and capable of expressing the binding domain fusion
proteins, the transcriptional activation domain fusion
proteins, and the Reporter Gene product(s) are provided
30 within the host cell, by mating of cells containing the
expression constructs, or by cell fusion, transformation,
electroporation, microinjection, etc. For example, GAL4 and
VP16 are functional in animal cells and thus the desired
binding or activation domain thereof can be used in, e.g.,
35 yeast or mammalian cells. In a specific embodiment in which
the assay is carried out in mammalian cells (e.g., hamster
cells), the DNA binding domain is the GAL4 DNA binding

- 23 -

domain, the activation domain is the herpes simplex virus
VP16 transcriptional activation domain, and the Reporter Gene
contains the desired coding sequence operably linked to a
minimal promoter element from the adenovirus E1B gene driven
5 by several GAL4 DNA binding sites (see Fearon et al., 1992,
Proc. Natl. Acad. Sci. USA 89:7958-7962). As will be
apparent, other DNA binding domains, activation domains,
promoters, and/or DNA binding sites can be used, as long as
the DNA binding sites are recognized by the DNA binding
10 domains, and the promoter is operative in the cells chosen in
which to carry out the assay of the invention. The host cell
used should not express an endogenous transcription factor
that binds to the same DNA site as that recognized by the DNA
binding domain fusion population. Also, preferably, the host
15 cell is mutant or otherwise lacking an endogenous, functional
form of the Reporter Gene(s) used in the assay.

        In a specific embodiment, transcription of the
Reporter Gene is detected by a linked replication assay. For
example, as described by Vasavada et al. (1991, Proc. Natl.
20 Acad. Sci. USA 88:10686-10690), for use in animal cells, a
Reporter Gene under the control of the E1B promoter, which
promoter in turn is controlled by GAL4 DNA binding sites,
encodes the SV40 T antigen. In the presence of reconstituted
GAL4 DNA binding domain-activation domain (composed of two
25 interacting fusion proteins), SV40 T antigen is produced from
the Reporter Gene. If a plasmid is present that contains the
SV40 origin of replication, this plasmid will replicate only
upon the production of SV40 T antigen. Thus, replication of
such a plasmid is used as an indicator of protein-protein
30 interaction. Constructing one or both of the plasmids
encoding the fusion proteins of the assay to contain an SV40
origin of replication means that replication of these
plasmids will be an indication of Reporter Gene activity.
Sensitivity to DpnI can be used to destroy unreplicated
35 plasmids according to the methods described in Vasavada et
al. (1991, Proc. Natl. Acad. Sci. USA 88:10686-10690). In an

alternative embodiment, alternatively to an SV40 origin of
replication, a polyoma virus replicon can be employed (id.)

Preferably, the protein-protein interactions are
assayed according to the method of the invention in yeast
5 cells, e.g., Saccharomyces cerevisiae or Schizo-saccharomyces
pombe. Various vectors for producing the two fusion protein
populations and host strains for conducting the assay are
known and can be used (see, e.g., Fields et al., U.S. Patent
No. 5,468,614 dated November 21, 1995; Bartel et al., 1993,
10 "Using the two-hybrid system to detect protein-protein
interactions," in Cellular Interactions in Development,
Hartley, D.A. (ed.), Practical Approach Series xviii, IRL
Press at Oxford University Press, New York, NY, pp. 153-179;
Fields and Sternglanz, 1994, TIG 10:286-292). By way of
15 example but not limitation, yeast strains or derivative
strains made therefrom which can be used are (see Section 6.3
and its subsections) N105, N106, N105', N106', and YULH; the
respective genotypes of these strains are set forth in
Section 6.3, infra. Exemplary strains that can be modified
20 to create reporter strains (containing the desired Reporter
Gene for use in the assay of the invention) also include the
following:

Y190: MATa, ura3-52, his3-200, lys2-801, ade2-101, trp1-901,
       leu2-3,112, gal4Δ, gal80Δ, cyh^{r}2, LYS2::GAL1$_{UAS}$-HIS3$_{TATA}$-
       HIS3, URA3::GAL1$_{UAS}$-GAL1$_{TATA}$-lacZ (available from Clontech,
25   Palo Alto, CA; Harper et al., 1993, Cell 75:805-816).
       Y190 contains HIS3 and lacZ Reporter Genes driven by
       GAL4 binding sites.

CG-1945: MATa, ura3-52, his3-200, lys2-801, ade2-101,
       trp1-901, leu2-3,112, gal4-542, gal80-538, cyh^{r}2,
       LYS2::GAL1$_{UAS}$-GAL1$_{TATA}$-HIS3, URA3::GAL1$_{UAS\ 17\ mers}$
30     $_{(x3)}$-CYC1$_{TATA}$-lacZ (available from Clontech). CG-1945
       contains HIS3 and lacZ Reporter Genes driven by GAL4
       binding sites.

Y187: MATα, ura3-52, his3-200, ade2-101, trp1-901,
       leu2-3,112, gal4Δ, gal80Δ, URA3::GAL1$_{UAS}$-GAL1$_{TATA}$-lacZ
       (available from Clontech). Y187 contains a lacZ
       Reporter Gene driven by GAL4 binding sites.
35
SFY526: MATa, ura3-52, his3-200, lys2-801, ade2-101, trp1-
       901, leu2-3,112, gal4-542, gal80-538, can^r,
       URA3::GAL1-lacZ (available from Clontech). SFY526

- 25 -

contains *HIS3* and *lacZ* Reporter Genes driven by GAL4 binding sites.

HF7c: *MATa, ura3-52, his3-200, lys2-801, ade2-101, trp1-901, leu2-3,112, gal4-542, gal80-538, LYS2::GAL1-HIS3, URA3::GAL1$_{UAS\ 17\ MERS\ (x3)}$-CYC1-lacZ* (available from Clontech). HF7c contains *HIS3* and *lacZ* Reporter Genes driven by GAL4 binding sites.

YRG-2: *MATα, ura3-52, his3-200, lys2-801, ade2-101, trp1-901, leu2-3,112, gal4-542, gal80-538, LYS2::GAL1$_{UAS}$-GAL1$_{TATA}$-HIS3, URA3::GAL1$_{UAS\ 17\ mers\ (x3)}$-CYC1-lacZ* (available from Stratagene). YRG-2 contains *HIS3* and *lacZ* Reporter Genes driven by GAL4 binding sites.

Many other strains commonly known and available in the art can be used.

Consistent with convention in the art, wild-type gene names throughout this application are all capitalized and italicized; mutant gene names are lower case and italicized -- except for *lacZ* for which the functional, non-mutant gene is written lower case, italicized.

If not already lacking in endogenous Reporter Gene activity, cells mutant in the Reporter Gene may be selected by known methods, or the cells can be made mutant in the target Reporter Gene by known gene-disruption methods prior to introducing the Reporter Gene (Rothstein, 1983, Meth. Enzymol. 101:202-211).

In a specific embodiment, plasmids encoding the different fusion protein populations can be both introduced into a single host cell (*e.g.*, a haploid yeast cell) containing one or more Reporter Genes, by cotransformation, to conduct the assay for protein-protein interactions. As a preferred alternative to cotransformation of expression constructs, mating (*e.g.*, of yeast cells) or cell fusion (*e.g.*, of mammalian cells) can be employed for delivery of a binding domain fusion expression construct and an activation domain fusion expression construct into a single cell. In a mating-type assay, conjugation of haploid yeast cells of opposite mating type that have been transformed with a binding domain fusion expression construct (preferably a

plasmid) and an activation (or inhibitor) domain fusion
expression construct (preferably a plasmid), respectively,
delivers both constructs into the same diploid cell. The
mating type of a strain may be manipulated as desired, by
5 transformation with the HO gene (Herskowitz and Jensen, 1991,
Meth. Enzymol. 194:132-146).

         In a preferred embodiment, a yeast interaction
mating assay is employed, using two different types of host
cells, strain-types a and α, of the yeast *Saccharomyces*
10 *cerevisiae* (Figure 2). The host cell preferably contains at
least two Reporter Genes, containing a binding site for the
DNA-binding domain (*e.g.*, of a transcriptional activator),
such that the Reporter Gene is transcriptionally activated
when the DNA-binding domain is in proximity to an activator
15 domain of a transcriptional activator. The activator domain
and DNA binding domain are each parts of chimeric proteins
formed from the two respective populations of proteins.

         One type of host cell, for example the a strain,
hosts a library of chimeric genes that encode hybrid proteins
20 that are all fusions of different nucleotide sequences (*e.g.*,
gene sequences) to the DNA-binding domain of a
transcriptional activator like GAL4 (see by way of example
Section 6.1.7). These hybrid proteins are capable of
recognizing the DNA-binding site on the Reporter Gene. In a
25 preferred embodiment of this invention, the library of DNA-
binding domain chimeric genes is introduced into the host
cell as a set of plasmids. These plasmids are preferably
capable of autonomous replication in a host yeast cell and
preferably can also be propagated in *E. coli*. The plasmid
30 contains a promoter directing the transcription of the DNA
binding domain fusion gene, and a transcriptional termination
signal. The plasmid preferably also contains a selectable
marker gene, the expression of which in the host cell permits
selection of cells containing the marker gene from cells that
35 do not contain the selectable marker, upon incubation of the
cells in an environment in which substantial death of the
cells occurs in the absence of expression of the selectable

- 27 -

marker.   The plasmid can be single-copy or multi-copy.
Single-copy yeast plasmids that have the yeast centromere in
them may also be used to express the activation and DNA-
binding domain fusions (Elledge et al., 1988, Gene
5 70:303-312).   In another embodiment of the invention, the
DNA-binding chimeric genes are introduced directly into the
yeast chromosome via homologous recombination.   The
homologous recombination for these purposes is mediated
through yeast sequences that are not essential for vegetative
10 growth of yeast, e.g., *MER2*, *MER1*, *ZIP1*, *REC102*, or *ME14*
gene.

In yet another embodiment of the invention,
alternatively to plasmids, bacteriophage vectors such as λ
vectors are used as the DNA binding domain vectors and/or
15 activation domain vectors to make, e.g., the respective cDNA
libraries.   The use of λ vectors generally makes it faster
and easier to generate such libraries than with the use of
plasmid vectors.

The second type of yeast host, for example the
20 strain α, hosts a library of chimeric genes encoding hybrid
proteins that are all fusions of different genes to the
activation domain of a transcriptional activator (see by way
of example Section 6.1.7).   Preferably, this library is
plasmid-borne, and the plasmids are capable of replication in
25 both *E. coli* and yeast.   The plasmid contains a promoter
directing the transcription of the activation domain fusion
gene, and a transcriptional termination signal.   The plasmid
preferably also contains a selectable marker gene, the
expression of which in the host cell permits selection of
30 cells containing the marker gene from cells that do not
contain the selectable marker.   In another embodiment of the
invention the DNA-binding chimeric genes are introduced
directly into the yeast chromosome via homologous
recombination.   The homologous recombination for these
35 purposes is mediated through yeast sequences that are not
essential for vegetative growth of yeast.

In one embodiment of the invention, the DNA-binding
domain and the activation domain arise from the same
transcriptional activator where these functions reside in
separate domains. In another embodiment, the DNA-binding and
5 the activation domains may be from different transcriptional
activators. Preferably, the two chimeric gene libraries are
made from cDNA from various sources, for example, different
human tissues, fused to the DNA-binding or the activation
domains, respectively (see by way of example Section 6.1.6).
10 These cDNA libraries may be derived from subtracted or
normalized cDNA populations. In other specific embodiments,
the fusions are of genomic, synthetic, viral or bacterial DNA
fused to the DNA-binding domains or the activation domains of
the transcriptional activator.

15   In a specific embodiment, the invention provides a
method of detecting one or more protein-protein interactions
comprising (a) recombinantly expressing in a first population
of yeast cells of a first mating type, a first population of
first fusion proteins, each first fusion protein comprising a
20 first protein sequence and a DNA binding domain, in which the
DNA binding domain is the same in each said first fusion
protein; wherein said first population of yeast cells
contains a first nucleotide sequence operably linked to a
promoter driven by one or more DNA binding sites recognized
25 by said DNA binding domain such that an interaction of a
first fusion protein with a second fusion protein, said
second fusion protein comprising a transcriptional activation
domain, results in increased transcription of said first
nucleotide sequence, and in which said first population of
30 first fusion proteins has a complexity of at least 1,000; (b)
negatively selecting to eliminate those yeast cells
expressing said first population of first fusion proteins in
which said increased transcription of said first nucleotide
sequence occurs in the absence of said second fusion protein;
35 (c) recombinantly expressing in a second population of yeast
cells of a second mating type different from said first
mating type, a second population of said second fusion

proteins, each second fusion protein comprising a second
protein sequence and an activation domain of a
transcriptional activator, in which the activation domain is
the same in each said second fusion protein, and in which `

5 said second population of second fusion proteins has a
complexity of at least 1,000; (d) mating said first
population of yeast cells with said second population of
yeast cells to form a population of diploid yeast cells,
wherein said population of diploid yeast cells contains a

10 second nucleotide sequence operably linked to a promoter
driven by a DNA binding site recognized by said DNA binding
domain such that an interaction of a first fusion protein
with a second fusion protein results in increased
transcription of said second nucleotide sequence, in which

15 the first and second nucleotide sequences can be the same or
different; and (e) detecting said increased transcription of
said first and/or second nucleotide sequence, thereby
detecting an interaction between a first fusion protein and a
second fusion protein.

20   In a preferred embodiment, the two libraries of
chimeric genes are combined by mating the two yeast strains
on solid media for a period of approximately 6-8 hours (see
Section 6.1.1). In a less preferred embodiment, the mating
is performed in liquid media. The resulting diploids contain

25 both the kinds of chimeric genes, i.e., the DNA-binding
domain fusion and the activation domain fusion. The
interaction between the two hybrid proteins within a diploid
cell causes the activation domain to be in close proximity to
the DNA-binding domain of the transcriptional activator.

30 This in turn causes reconstitution of the transcriptional
activator and is monitored by the activity of the Reporter
Gene. Thus, when two libraries M and N are mated together,
an M x N screen for interacting proteins is performed.

          In a preferred embodiment, the two host strains are

35 preferably of the mating type a and α of the yeast
Saccharomyces cerevisiae. Each mating type of the host
preferably has at least two Reporter Genes that each contain

- 30 -

one or more recognition sites for the DNA-binding domain.
Preferably, the Reporter Gene(s) are the *URA3*, *HIS3* and/or
the *lacZ* (see, *e.g.*, Rose and Botstein, 1983, Meth. Enzymol.
101:167-180) gene that have been manipulated so as to contain

5 recognition sites (preferably at least two) in the promoter
for the DNA-binding domain of GAL4 (see by way of example
Section 6.3.5) (Figure 2). In other embodiments, Reporter
Genes comprising the functional coding sequences of genes,
including but not limited to, Green Fluorescent Protein (*GFP*)

10 (Cubitt et al., 1995, Trends Biochem. Sci. 20:448-455),
luciferase, *LEU2*, *LYS2*, *ADE2*, *TRP1*, *CAN1*, *CYH2*, *GUS*, *CUP1*
(encoding metallothionein which confers resistance to copper)
or chloramphenicol acetyl transferase (CAT) may be used,
operatively linked to a promoter driven by DNA binding

15 site(s) recognized by the DNA binding domain being employed
in the assay to form a fusion population. *LEU2*, *LYS2*, *ADE2*
and *TRP1* are selectable markers, *i.e.*, their activity results
in prototrophic growth in media lacking the nutrients encoded
by these genes, while the activity of luciferase, *GUS* and *CAT*

20 are preferably monitored enzymatically. Preferably, *CAN1* and
CYH2 Reporter Genes are used to carry out negative selection
in the presence of canavanine and cyloheximide, respectively
(see *infra*), rather than to detect an interacting pair of
proteins. With respect to GFP, the natural fluorescence of

25 the protein is detected. In another embodiment, the
expression of Reporter Genes that encode proteins can be
detected by immunoassay, *i.e.*, by detecting the
immunospecific binding of an antibody to such protein, which
antibody can be labeled, or alternatively, which antibody can

30 be incubated with a labeled binding partner to the antibody,
so as to yield a detectable signal. Alam and Cook (1990,
Anal. Biochem. 188:245-254) disclose non-limiting examples of
detectable marker genes that can be constructed so as to be
operably linked to a transcriptional regulatory region

35 responsive to a reconstituted transcriptional activator used
in the method of the invention, and thus used as Reporter
Genes. As will be apparent, use of a particular Reporter

- 31 -

Gene should be conducted in cells mutant or otherwise lacking
in functional versions of the Reporter Gene. Thus, for
example, for (positive or negative) selection for *URA3*
Reporter Gene activity, the host cell should be homozygous
5 mutant (point mutation or deleted or otherwise lacking
function of the gene in both alleles) so as to lack
endogenous *URA3* activity. Similarly, in the use of a *LYS2*
Reporter Gene, the host cell should be homozygous mutant for
LYS2, in the use of a *CAN1* Reporter Gene for negative
10 selection, the host cell should be homozygous mutant for
CAN1, in the use of a *CYH2* Reporter Gene for negative
selection, the host cell should be homozygous mutant for
CYH2, etc., in cases in which the host cell has an endogenous
form of the Reporter Gene.

15   The activation of Reporter Genes like *URA3* or *HIS3*
enables the cells to grow in the absence of uracil or
histidine, respectively, and hence serves as a selectable
marker. Thus, after mating, the cells exhibiting protein-
protein interactions are selected by their abilities to grow
20 in media lacking the requisite ingredient like uracil or
histidine, respectively (referred to as -URA (minus URA) and
-HIS medium, respectively) (see by way of example Section
6.3-6.5). In a specific embodiment, -HIS medium preferably
contains 3-amino-1,2,4-triazole (3-AT), which is a
25 competitive inhibitor of the *HIS3* gene product and thus
requires higher levels of transcription in the selection (see
Durfee et al., 1993, Genes Dev. 7:555-569). Similarly,
6-azauracil, which is an inhibitor of the *URA3* gene product,
can be included in -URA medium (Le Douarin et al., 1995,
30 Nucl. Acids Res. 23:876-878). Alternatively to detecting
URA3 gene activity by selecting in -URA medium, *URA3* gene
activity can be detected and/or measured by determining the
activity of its gene product, orotidine-5'-monophosphate
decarboxylase (Pierrat et al., 1992, Gene 119:237-245;
35 Wolcott et al., 1966, Biochem. Biophys. Acta 122:532-534).
In other embodiments of the invention, the activities of the
reporter genes like *lacZ* or GFP are monitored by measuring a

detectable signal (e.g., fluorescent or chromogenic) that
results from the activation of these Reporter Genes. For
example, *lacZ* transcription can be monitored by incubation in
the presence of a chromogenic substrate, such as X-gal

5 (5-bromo-4-chloro-3-indolyl-β-D-galactoside), for its encoded
enzyme, β-galactosidase. The pool of all interacting
proteins isolated by this manner from mating the two
libraries is termed the "interactive population" (see by way
of example Figure 3).

10   In a preferred embodiment of the invention, false
positives arising from transcriptional activation by the DNA
binding domain fusion proteins in the absence of a
transcriptional activator domain fusion protein are prevented
or reduced by negative selection for such activation within a

15 host cell containing the DNA binding fusion population, prior
to exposure to the activation domain fusion population. By
way of example, if such cell contains *URA3* as a Reporter
Gene, negative selection is carried out by incubating the
cell in the presence of 5-fluoroorotic acid (5-FOA, which

20 kills URA+ cells (Rothstein, 1983, Meth. Enzymol.
101:167-180). Hence, if the DNA-binding domain fusions by
themselves activate transcription, the metabolism of 5-FOA
will lead to cell death and the removal of self-activating
DNA-binding domain hybrids. By way of another example, if

25 *LYS2* is present as a Reporter Gene in the cell, negative
selection is carried out by incubating the cell in the
presence of α-amino-adipate (Chatoo et al., 1979, Genetics
93:51), which kills LYS⁺ cells. In another embodiment, if
CAN1 is present as a Reporter Gene in the cell, negative

30 selection is carried out by incubating the cell in the
presence of canavanine (*CAN1* encodes an arginine permease
that renders the cell sensitive to the lethal effects of
canavanine) (Sikorski et al., 1991, Meth. Enzymol.
194:302-318). In yet another embodiment, if *CYH2* is present

35 as a Reporter Gene in the cell, negative selection is carried
out by incubating the cell in the presence of cycloheximide
(*CYH2* encodes the L29 protein of the yeast ribosome; the

wild-type L29 protein is sensitive to cycloheximide which
thus blocks protein synthesis, resulting in cell death)
(Sikorski et al., 1991, Meth. Enzymol. 194:302-318).   Such
negative selection with the DNA-binding domain fusion
5 population helps to avoid false positives that become
amplified through the preferred processing steps of the
invention, and which becomes more troublesome as the
complexity of the assayed populations increases.   In another
embodiment, the DNA-binding domain fusion population can be
10 subjected to negative immunoselection by use of antibodies
specific to the expressed protein product of a Reporter Gene;
in this embodiment, cells expressing a protein that is
recognized by the antibody are removed and the fusion
constructs from the remaining cells are kept for use in the
15 interaction assay.   In yet another embodiment, negative
selection can be carried out by plating the DNA-binding
domain fusion population on medium selective for interaction
(e.g., minus URA or minus HIS medium if the Reporter Gene is
URA3 or HIS3, respectively), following which all the
20 surviving colonies are physically removed and discarded.
Negative selection involving the use of a selectable marker
as a Reporter Gene and the presence in the cell medium of an
agent toxic or growth inhibitory to the host cells in the
absence of Reporter Gene transcription is preferred, since it
25 allows high throughput, i.e., a much greater number of cells
to be processed much more easily than alternative methods.

As will be apparent, negative selection can also be
carried out on the activation domain fusion population prior
to interaction with the DNA binding domain fusion population,
30 by similar methods, alone or in addition to negative
selection of the DNA binding fusion population.

In another embodiment, negative selection can also
be carried out on the recovered pairs of protein
interactants, by known methods (see, e.g., Bartel et al.,
35 1993, BioTechniques 14(6):920-924) although pre-negative
selection (prior to the interaction assay), as described
above, is preferred.   For example, each plasmid encoding a

- 34 -

protein (peptide or polypeptide) fused to the activation
domain (one-half of a detected interacting pair) can be
transformed back into the original screening strain, either
without any other plasmid, or with a plasmid encoding only
5 the DNA-binding domain, the DNA-binding domain fusion to the
detected interacting protein (the second half of the detected
interacting pair), or the DNA-binding domain fusion to an
irrelevant protein; a positive interaction detected with any
plasmid other than that encoding the DNA-binding domain
10 fusion to the detected interacting protein is deemed a false
positive and eliminated from further use.

         In a preferred embodiment of the invention, the
DNA-binding domain library is introduced into a host strain
that has *URA3* as a reporter gene. This library should not
15 activate transcription by itself. To weed out DNA-binding
domain fusions that activate transcription by themselves
(carry out negative selection), the yeast transformants
containing the DNA-binding domain library are plated out on
media that contain the chemical 5-fluoroorotic acid (5-FOA).
20 In order to easily detect the protein-protein interactions
between proteins in complex populations as provided by the
methods of the present invention, it is preferred to use a
host cell containing at least two, preferably three, Reporter
Genes (*e.g.*, *HIS3, URA3, lacZ* operably linked to a DNA
25 binding site of a transcription activator that is recognized
by the DNA binding domain part of the fusion protein, in a
yeast host cell), and to carry out negative selection among
the DNA binding domain-fusion protein population (*e.g.*, by
use of 5-FOA and a *URA3* Reporter Gene); and to use a yeast
30 mating assay in which the mating is performed on a solid
phase, which increases the percentage of productive mating
events that can be recovered.

         In a specific embodiment, a DNA binding domain
fusion library is expressed from a first plasmid population,
35 and a transcription activation domain fusion library is
expressed from a second plasmid population, and each plasmid
contains a selectable marker. For example, the first plasmid

- 35 -

population can express *TRP1*, and the second plasmid
population can express *LEU2*, or some other gene encoding an
essential amino acid so that the presence of the plasmid can
be selected for in medium lacking the amino acid. In a
5 preferred embodiment, the first plasmid population is
expressed in a yeast strain of a first mating type (selected
from between a and α), and which yeast strain is deficient in
endogenous *URA3* and *HIS3*, and contains *URA3* as a Reporter
Gene and optionally also *lacZ* as a Reporter Gene. In a
10 preferred embodiment, the second plasmid population is
expressed in a yeast strain of a second mating type different
from the first mating type, which yeast strain is deficient
in endogenous *URA3* and *HIS3*, and contains *HIS3* as a Reporter
Gene and optionally also *lacZ* as a Reporter Gene. Yeast
15 cells of the first mating type are transformed with the first
plasmid population, and are positively selected for the
plasmids and are negatively selected for false positive
transcriptional activation by incubating the cells in an
environment (e.g., liquid medium, and/or solid phase plates)
20 lacking the selectable marker (e.g., tryptophan) and
containing 5-FOA. Selected cells are pooled. Yeast cells of
the second mating type are transformed with the second
plasmid population, and are positively selected for the
plasmids by incubating the cells in an environment lacking
25 the appropriate selectable marker, e.g., leucine. Selected
cells are pooled. Both groups of pooled cells are mixed
together and mating is allowed to occur on a solid phase.
The resulting diploid cells are then transferred to selective
media, that selects for the presence of each plasmid and for
30 activation of Reporter Genes, *i.e.*, in this embodiment,
medium lacking uracil, histidine, tryptophan and leucine, and
optionally, also containing 3-amino-1,2,4-triazole.

In specific embodiments, the invention also
provides purified cells of a single yeast strain of mating
35 type a, that is mutant in endogenous *URA3* and *HIS3*, and
contains functional *URA3* coding sequences under the control
of a promoter containing GAL4 binding sites, and contains

- 36 -

functional *lacZ* coding sequences under the control of a
promoter containing GAL4 binding sites; and also provides
purified cells of a single yeast strain of mating type α,
that is mutant in endogenous *URA3* and *HIS3*, and contains
5 functional *URA3* coding sequences under the control of a
promoter containing GAL4 binding sites, and contains
functional *lacZ* coding sequences under the control of a
promoter containing GAL4 binding sites. A kit is also
provided, comprising in one or more containers cells of the
10 foregoing strains. In a specific embodiment, the kit further
comprises in one or more containers (a) a first vector
comprising (i) a promoter; (ii) a nucleotide sequence
encoding a DNA binding domain, operably linked to the
promoter; (iii) means for inserting a DNA sequence encoding a
15 protein into the vector in such a manner that the protein is
capable of being expressed as part of a fusion protein
containing the DNA binding domain; (iv) a transcription
termination signal operably linked to the nucleotide
sequence; (v) a means for replicating in the cells of the
20 above-described yeast strains; and (c) a second vector
comprising (i) a promoter; (ii) a nucleotide sequence
encoding an activation domain of a transcriptional activator,
operably linked to the promoter; (iii) means for inserting a
DNA sequence encoding a protein into the vector in such a
25 manner that the protein is capable of being expressed as part
of a fusion protein containing the activation domain of a
transcriptional activator; (iv) a transcription termination
signal operably linked to the nucleotide sequence; and (v) a
means for replicating in the cells of the above-described
30 yeast strains. The means for inserting a DNA sequence can be
one or more restriction endonuclease recognition sites
suitably located within the vector.

In a preferred embodiment of the invention, after
an interactive population is obtained, the DNA sequences
35 encoding the pairs of interactive proteins are isolated by a
method wherein either the DNA-binding domain hybrids or the
activation domain hybrids are amplified specifically in an

- 37 -

individual reaction (see by way of example Section 6.9).
Preferably, both the DNA-binding fusion sequences and the
activation domain fusion sequences are amplified, in separate
respective reactions. Preferably, the amplification is
5 carried out by polymerase chain reaction (PCR) (U.S. Patent
Nos. 4,683,202. 4,683,195 and 4,889,818; Gyllenstein et al.,
1988, Proc. Natl. Acad. Sci. USA 85:7652-7656; Ochman et al.,
1988, Genetics 120:621-623; Loh et al., 1989, Science
243:217-220; Innis et al., 1990, PCR Protocols, Academic
10 Press, Inc., San Diego, CA), using pairs of oligonucleotide
primers that are specific to either the DNA-binding domain
hybrids or the activation domain hybrids in the PCR reaction
(see by way of example Section 6.1.8). This PCR reaction can
also be performed on pooled cells expressing interacting
15 protein pairs, preferably pooled arrays of interactants.
Other amplification methods known in the art can be used,
including but not limited to ligase chain reaction (see EP
320,308) use of Qβ replicase, or methods listed in Kricka et
al., 1995, Molecular Probing, Blotting, and Sequencing, chap.
20 1 and table IX, Academic Press, New York.

      In another embodiment of the invention, the
plasmids encoding the DNA-binding domain hybrid and the
activation domain hybrid proteins are isolated from yeast
cells by transforming the yeast DNA into *E. coli* and
25 recovering the plasmids from *E. coli* (see e.*g*., Hoffman et
al., 1987, Gene 57:267-272). This is possible when the
plasmid vectors used for both the DNA-binding domain and the
activation domain hybrids are shuttle vectors that can
replicate both in *E. coli* and in yeast. Many such shuttle
30 vectors are known in the art and can be used. Alternatively,
if a shuttle vector is not used, the yeast vector can be
isolated, and the insert encoding the fusion protein
subcloned into a bacterial expression vector for growth in
bacteria. Growing up the interacting clones in bacteria
35 yields large quantities without the use of amplification
reactions such as PCR.

5.2. CHARACTERIZATION OF INTERACTIVE POPULATIONS THAT ARE DIFFERENTIALLY EXPRESSED BY A PARTICULAR TISSUE TYPE, DISEASE STATE OR STAGE OF DEVELOPMENT, AND CREATION OF "PROTEIN INTERACTION MAPS".

5   An important object of the present invention is to provide a method to identify protein-protein interactions that are unique to particular disease states, stages of development, or tissue type. An analysis of the interacting proteins between two populations of proteins ("M x N

10   analysis") performed in parallel on two types of tissue or disease states, wherein both the M and N populations are preferably identical and are derived from the same type of tissue or disease state, will yield the respective interactive protein populations for each type. The

15   differences between the two interactive populations will yield the protein-protein interactions that are characteristic of or unique to a particular tissue type or disease state. Hence, it is desired to identify and isolate the protein-protein interactions that are unique to a complex

20   population. This is preferably achieved by coding, pooling and arraying strategies for the interactants as described below and deconvolution of the arrayed interactants by sequencing a Quantitative Expression Analysis (QEA™ method), SEQ-QEA™ method, and/or other methods that facilitate

25   analysis of the interactants (e.g., SAGE (Velculescu et al., 1995, Science 270:484-487). Alternatively, sequencing of individual interactants provides a method for identifying the interacting genes that does not necessarily use pooling or require deconvolution. Thus, in this alternative embodiment,

30   clones of interactants can be recovered, e.g., from the interactant-positive yeast cells, amplified or grown up in bacteria, and subjected to sequence analysis. Sequencing can be carried out by any of numerous methods known in the art (see e.g., Sanger et al., 1977, Proc. Natl. Acad. Sci. USA

35   74(12):5463-5467). In a specific embodiment, to enhance throughput, a multiplex sequencing analysis can be conducted. For example, in a multiplex sequencing analysis, one can

- 39 -

carry out dideoxy sequencing reactions with just one of the
dideoxynucleotides, e.g., ddT, using a different dye on the
dideoxynucleotide in the reaction with DNA of each of four
separate interactant pairs, which reaction products are then
5 pooled together and subjected to electrophoresis. Comparing
the pattern of bands formed by DNA of interactant pairs from
different populations identifies differences, indicating an
interacting protein specific to that population. The DNA for
such a protein can then be sequenced fully. Moreover,
10 identical patterns of bands for a single dye between pooled
groups identifies interactions which share the same partners,
thus saving sequencing DNA encoding a common interacting
protein over and over again. This method would raise
throughput four-fold.
15

5.2.1. DETERMINATION OF ALL THE DETECTABLE PROTEIN-PROTEIN INTERACTIONS

Cells containing interacting protein pairs are
identified as described above, by detecting Reporter Gene
20 expression. Determining all the detectable pairs of
interactions then employs pooling and two sets of
deconvolution reactions. The first set characterizes all the
"M" interacting partners; the second set characterizes the
"N" interacting partners. Preferably, DNA of cells
25 containing interacting proteins is subjected to an
amplification reaction that specifically amplifies the DNA-
binding fusion sequences and, in a separate reaction, the
activation domain fusion sequences. In a preferred
embodiment, the characterizations of interacting partners are
30 performed by "the SEQ-QEA™ method" (see infra) on PCR
products that were generated with "M" or "N" specific
amplification primers, respectively (see by way of example
Figure 3). The "M"-specific amplification primers hybridize
specifically to and amplify sequences from one type of fusion
35 construct, e.g., the DNA binding fusion construct (e.g., by
hybridization to vector sequences flanking the inserted
variant protein coding sequences of population M that are

- 40 -

fused to the DNA binding domain sequences). The "N"-specific
amplification primers hybridize specifically to and amplify
sequences from the other type of fusion construct, the
activation domain fusion construct (e.g., by hybridization to
5 vector sequences flanking the inserted variant protein coding
sequence of population N that are fused to the activation
domain sequences). The PCR is preferably performed wherein
DNA-binding and activation domain fusion specific primers are
used to amplify the genes encoding the two interacting
10 proteins directly from yeast (see by way of example Section
6.1.8). This PCR product serves as a reservoir for further
analysis, including the QEA™ method, the SEQ-QEA™ method (see
infra) and sequencing, that leads to the identification of
interacting proteins, in particular, those that are
15 differentially expressed (e.g., stage-specific). The primers
used in the PCR reaction may be labelled, e.g., by
biotinylation or addition of fluorescent tags and may also
serve to introduce specific restriction endonuclease sites.
The labels are useful tools in the subsequent QEA™ method and
20 sequencing.

        Thus, in a specific embodiment, DNA isolated from
each cell containing each individual pair of interactants is,
in separate reactions, subjected to PCR to amplify the DNA
encoding the DNA-binding domain fusion protein, and DNA
25 encoding the activation domain fusion protein, respectively.
The DNA encoding the DNA-binding domain fusion protein and
the DNA encoding the activation domain fusion protein are
each subjected to sequencing analysis to determine its
sequence and thus the sequence of the interacting protein
30 that formed a part of the fusion protein. In this manner,
each individual pair of interactants is identified.
Alternative methods that can be used to identify individual
pairs of interactants are described in Sections 5.2.2 to
5.2.6.2.
35

5.2.2. CLASSIFICATION OF THE ARRAYED POOLS OF INTERACTANTS BY THE QEA™ METHOD AND THE SEQ-QEA™ METHOD

A Quantitative Expression Analysis method (QEA™ method) produces signals comprising target subsequence presence and a representation of the length in base pairs along a nucleic acid between adjacent target subsequences by measuring the results of recognition reactions on DNA (e.g., cDNA or genomic DNA) mixtures. A QEA™ method provides an economical, quantitative, and precise classification of DNA sequences, either in arrays of single sequence clones or in mixtures of sequences, without actually sequencing the DNA. Preferably, all the signals taken together have sufficient discrimination and resolution so that each particular DNA sequence in a sample may be individually classified by the particular signals it generates, and with reference to a database of DNA sequences possible in the sample individually determined. These signals are preferably optical, generated by fluorochrome labels and detected by automated optical detection technologies. The signals are generated by detecting the presence or absence of short DNA subsequences within a nucleic acid sequence of the sample analyzed. The subsequences are detected by use of recognition means, or probes for the subsequences. A detailed description of the QEA™ methods is provided in the U.S. patent applications Serial No. 08/547,214 filed on October 24, 1995, and Serial No. to be assigned, filed on even date herewith, both by Rothberg et al. and entitled "Method and apparatus for classifying, identifying, or quantifying DNA sequences in a sample without sequencing", which are incorporated by reference herein in their entireties. QEA™ methods that can be used are also described in Section 5.4, infra, and, by way of example, in Section 6.1.12.

A QEA™ method reveals the distribution (both qualitative and quantitative) of genes within a population. Thus when comparing two interactive populations to which a QEA™ method is applied, the differential presence of genes

- 42 -

between two interactive populations is identified as unique
or increased or decreased intensity bands after size
separation such as in a denaturing polyacrylamide gel (see by
way of example Section 6.1.12). In a preferred embodiment of
5 the invention, the identity of the gene producing each band
is determined by a modification of the QEA™ method called the
SEQ-QEA™ method (see by way of example Sections 6.1.12.2 and
6.1.12.5). The SEQ-QEA™ method (for Sequencing QEA™)
provides a method to identify the 4 terminal nucleotides next
10 to a subsequence that was used as a recognition site in the
QEA™ method. Thus, by combining the information from the
QEA™ method and the SEQ-QEA™ method, it is possible to
classify and identify precisely the DNA sequences present in
an interactive population without sequencing. A description
15 of SEQ-QEA™ methods is provided in Section 5.4.4 and, by way
of example, in Sections 6.1.12.2 (and its subsections) and
6.1.12.5.

5.2.3.   ARRAYING AND CODING STRATEGIES
20   FOR AN INTERACTIVE POPULATION

In a preferred embodiment, "interactive colonies"
are arrayed into wells on microtiter plates. "Interactive
colonies" are those colonies that emerge as a result of the
selection of interacting proteins. A deconvolution strategy
25 allows for a characterization of both members of each pair of
interacting proteins (from all the individual wells) without
sequencing each pair individually. In this way, the proteins
expressed in each well are characterized and statistics can
be gathered as to the frequency of the types of interactions.
30 We refer to the catalog of interacting proteins as a "protein
interaction map". This characterization can be further used
to identify the genes of interest directly or to indicate the
specific physical locations in the array of clones that
should be sequenced to determine (or confirm) the identities.
35 Thus, this process provides information on protein-protein
interactions characterizing a population of interest.

The differences between the patterns of protein interactions in different types of tissue (e.g., diseased versus normal, different stages of development, etc.) provide information that can be much more valuable than the knowledge

5 of the interactions in a single tissue alone. Similarly, expression levels (e.g., as determined by the QEA™ method) yield greater value when they can be correlated with fundamental differences in various tissue samples. A protein interaction map of any given tissue or cell type will contain

10 many non-biological or unimportant interactions. However, a comparison of the interactions taking place between a disease state, and a "normal" state will be very informative, as this process of comparison tends to eliminate the unimportant interactions. Identifying the genes encoding interacting

15 proteins may also provide information on the putative biological functions of the genes of interest, which will help assess which of the interactions detected are likely to take place physiologically (some interactions in the protein interaction map might be artifacts of the method) and be of

20 heightened interest. It can also be valuable to review the differences in protein interaction maps with the results of a QEA™ method or other method of analyzing expression levels (e.g., SAGE (Velculescu et al., 1995, Science 270:484-487; Northern analysis) performed on a cDNA population prior to

25 performing an interaction screen according to the invention. For instance, the appearance of a new interaction in diseased tissue that is not present in normal tissue can be correlated with the QEA™ method or SAGE or Northern analysis measurements of the expression levels of the genes involved

30 in the interaction. Upregulation or co-regulation of the genes would serve to corroborate the protein interaction maps.

5.2.4. MAINTAINING LINKAGE BETWEEN PAIRS OF INTERACTING PROTEINS

35

The most preferable QEA™ method on the amplified products derived from a pool of interactive colonies

identifies the interactions that take place in a sample and
identifies the differences between samples while retaining
the linkage between a specific gene and the corresponding
interacting partner gene of the interacting pair. If the
5 QEA™ method is done on the entire pooled, interacting
population and this is compared to another entire population,
the linkage between interacting partners in each individual
sample (i.e., from each individual colony containing a
separate interacting pair) is lost. Preferable pooling
10 strategies are coupled with deconvolution strategies that
maintain linkage between the interacting partners and allow
identification of the interactive colony that gives rise to
each set of interacting partners. The invention provides a
method of determining one or more characteristics of or the
15 identities of nucleic acids encoding an interacting pair of
proteins from among a population of cells containing a
multiplicity of different nucleic acids encoding different
pairs of interacting proteins, said method comprising (a)
designating each group of cells containing nucleic acids
20 encoding an identical pair of interacting proteins as one
point of a multidimensional array in which the intersection
of axes in each dimension uniquely identifies a single said
group; (b) pooling all groups along a simple axis to form a
plurality of pooled groups; (c) amplifying from a first
25 aliquot of each pooled group a plurality of first nucleic
acids, each first nucleic acid comprising a sequence encoding
a first protein that is one-half of a pair of interacting
proteins; (d) amplifying from a second aliquot of each pooled
group a plurality of second nucleic acids, each second
30 nucleic acid comprising a sequence encoding a second protein
that is the other half of the pair of interacting proteins;
(e) subjecting said first nucleic acids from each pooled
group to size separation; (f) subjecting said second nucleic
acids from each pooled group to size separation; (g)
35 identifying which at least one of said first nucleic acids
are present in samples of first nucleic acids from a pooled
group from each axes in each dimension, thereby indicating

- 45 -

that said at least one first nucleic acid is present in said
array in the group designated at the intersection of said
axes in each dimension; and (h) identifying which at least
one of said second nucleic acids are present in samples of a
5 second nucleic acid from a pooled group from axes in each
dimension, thereby indicating that the said at least one
second nucleic acid is present in said array in the group
designated at the intersection of said axes in each
dimension; in which the first and second nucleic acids that
10 are indicated to be present in said array in a group
designated at the same intersection are indicated to encode
interacting proteins. In preferred aspects, such a method is
applied to colonies of yeast cells, each colony containing
nucleic acids encoding a different pair or interacting
15 proteins identified according to a method of the invention.
Exemplary pooling and deconvolution strategies are described
below.

Pooling and deconvolution strategies can be
characterized by the dimensionality of the pooling array. We
20 assume that N distinct colonies containing interacting pairs
of proteins have been identified. Sequencing of each pair of
interactors individually corresponds formally to a 1-
dimensional strategy in which each pool draws from one of
the N samples. This yields N pools in total. In higher
25 dimensions, the number of pools required is

$$D \times N^{1/D},$$

where D is the number of dimensions. (This assumes a square
grid). The maximum number of genes in each pool is the
number of colonies contributing to each of the pools. Again
30 assuming a square grid, the maximum number is

$$(max \ genes/pool) = N^{(D-1)/D},$$

where N is the total number of colonies used in a D-
dimensional pooling strategy.

Increasing the dimensionality D reduces the total
35 number of pools but increases the total number of genes that
can be in each pool. It is preferable to choose the largest
value for D such that the genes in a pool can still be

- 46 -

identified.   Thus, the optimal pooling strategy, *i.e.*, the
preferred choice for D, depends on the number of individual
genes that can be identified in a single pool as well as on
the total number of interactive colonies.

5   In order to standardize the pooling and
deconvolution strategy, it can be preferable to use a 2-
dimensional pooling and deconvolution strategy exclusively.
If the size of the interactive population is in the hundreds,
then a simple two-dimensional pooling strategy suffices.

10   Further details of preferred pooling and
deconvolution strategies are provided below.   In a specific
embodiment, strategies are automated.

5.2.5.   POOLING STRATEGIES

15 *2-dimensional pools*

           In a preferred embodiment of a 2-dimensional
strategy, the interactive colonies are arrayed in a 12 x 8
grid representing 96 different interactive colonies (as shown
in Figure 4A).   The cells from the rows and columns are then

20 pooled together and amplification (preferably PCR) is
performed on the pools of interactants.   Two sets of
amplification (*e.g.*, PCR) reactions, one specific for one
kind of the fusion protein (or M) and the other specific for
the second kind of fusion protein (or N), are then performed.

25 If the total number of interactants is small (<20), then
electrophoretic separation (*e.g.*, by polyacrylamide or
agarose gel electrophoresis) of the amplified (*e.g.*, PCR)
products is generally sufficient to distinguish the
interactants from one another (see Figure 4A).   In that case,

30 comparison of the amplified products from each row and column
identifies the interactive colony from which the amplified
product originated.   That is, the presence of a band in both
a sample from a pooled row and a sample from a pooled column
indicates that the band is present in the interactive colony

35 present at the intersection of the row and colony.   A perfect
symmetry (the same PCR product in two rows and columns)
indicates either the same pair of interactants repeating or

- 47 -

two pairs of interactants that have insert DNAs of identical lengths.

When the number of interactants is greater than 20 and within a few hundred, a 2-dimensional strategy is still
5 sufficient. However, distinct inserts may have the same lengths and may not be separated to adequate resolution, for example, by electrophoresis of PCR products. Therefore, in a preferred embodiment, to aid in the deconvolution, the QEA™ method applied to cDNA populations is performed with a 4-mer
10 or 6-mer recognition subsequence. The length of the recognition subsequence is adjusted to provide a resolvable number of the QEA™ method bands. Because the size of the inserts in interactive populations tends to be in the range of 0.5 to 3 kb when using mammalian cDNA libraries as source
15 of the populations, the use of 6-mer subsequences can necessitate that a large number of reactions be performed in order to ensure that every insert DNA contains two such subsequences and thus has been included in the QEA™ method. The use of 4-mer recognition subsequences provides more
20 frequent cutting and can alleviate this problem. As 4-mer subsequence "hits" occur more frequently than with 6-mer subsequences, the probability of including each interactant in the QEA™ method increases. Furthermore, by limiting the number of interactants in a given pool to 10 to 15, the
25 number of "bands" or genes in a QEA™ method can be limited to about 40, and thus provide an easily analyzable QEA™ method readout that can be used to deconvolute the pools. Exemplary protocols for a QEA™ method that can be used are described in Section 6.1.12 and its subsections (particularly 6.1.12.2).
30 In a preferred embodiment, the addition of the SEQ-QEA™ method to the above analysis further refines the deconvolution process by imparting more information to each band (see, by way of example, Section 6.1.12.2). Furthermore, the SEQ-QEA™ method aids in uniquely identifying
35 the bands from the QEA™ method reaction. This often is not possible using a the 4-mer QEA™ method alone as the information from such a QEA™ method reaction is generally not

sufficient to uniquely identify genes within a eukaryotic cDNA population made from total mRNA. The ability to identify unambiguously the bands in each pool and those in common between pools is the desired outcome of deconvolution.

5 The methods of QEA™ method (preferably 4-mer), preferably in combination with the SEQ-QEA™ method, resolve the identity of the bands in each pool, thus identifying the proteins that appear in an interacting pair, and in common between pools without the need for sequencing of the bands. By such

10 methods, the identified bands that appear, or appear at increased level, after the interaction assay of the invention is carried out wherein a first cDNA population forms both N and M populations, compared to the bands that appear after the interaction assay is carried out with a second cDNA

15 population forming both N and M, identifies differentially expressed proteins between the first and second cDNA populations that mediate protein-protein interactions.

3-dimensional pools

20   In the case of large interactive populations, a 3-dimensional coding and pooling strategy (Figures 4B-4C) is used. In the illustrated example of Figure 4B-4C, a total of 32 pools are used: 12 (pooled columns, 8 x 12 wells each) + 8 (pooled rows, 144 wells)+ 12 (pooled plates, 96). Each

25 pool will have a maximum of 144 genes (Figures 4B-4C). The QEA™ method and SEQ-QEA™ method are performed on the PCR products derived from each pool (separately for the DNA-binding fusions, and the activation fusions, respectively), and the intersection of three pooling dimensions is used to

30 identify the gene at each location. The SEQ-QEA™ method based on 4-mer subsequences may not be easy to interpret due to the large number of bands (genes) in each pool. Therefore, it can be preferable to use a large number of less common subsequence pairs (6-mers instead of 4-mers) to

35 discriminate between all the genes present.

5.2.6.   ALTERNATIVE STRATEGIES TO
CHARACTERIZE INTERACTIVE POPULATIONS

5.2.6.1.   SEQUENCE-BASED STRATEGIES
TO IDENTIFY PAIRS OF
INTERACTING PROTEINS

5

An alternative strategy involving gene-specific PCR provides means to identify the pair of genes coding for each set of interacting proteins, as described hereinbelow. The QEA™ method performed on the interactive populations identifies 'difference' bands (bands that differentiate one

10 interactive population from the other). In a pooling strategy, in which different colonies are pooled together before the QEA™ method, it is preferable to have means to indicate which colony gave rise to each band. This section describes means for performing sequencing studies to identify

15 which colony gives rise to each band. The methods in this section are based on sequencing, which also provides the identity of the sequence generating each QEA™ method band in question, the same sequences that encode the proteins responsible for the interactions.

20

A QEA™ method band includes knowledge of specific sub-sequences (which the recognition means, used in the QEA™ method reaction, detect). Specific PCR primers are designed based on these sub-sequences so as to be able to hybridize to and thus amplify only those bands in a pooled population that

25 contain these sub-sequences. Thus, these PCR primers are used to screen by PCR the entire interactive population. This is done by performing PCR with gene-specific primers, preferably on the original stored PCR products (both the DNA-binding domain-specific and activation domain-specific PCR

30 products), when pooled according to the two-dimensional or three-dimensional pooling strategies described above. A specific PCR product will be observed only if the particular PCR pool contains the gene that gives rise to the QEA™ method band. Deconvolution strategies can be carried out as

35 described above. Thus, e.g., a PCR product appearing at the intersection of a pooled row and pooled column (or pooled

plate, in a three-dimensional strategy) indicates that such PCR product arose from the colony situated at such intersection, and indicates that such PCR product contains the subsequences to which the primers were designed to
5 hybridize. By this method, the original mating pair that gives rise to the QEA™ method band can be identified and the sequence of the two genes that encode the interacting proteins can be confirmed by sequencing the respective DNA-binding domain and activation domain plasmids after isolating
10 these plasmids from the relevant colony.

5.2.6.2. <u>CREATION OF INTERACTIVE-GRIDS</u>

As a variation of the PCR-based strategy, a hybridization-based strategy can also be used to identify
15 interacting proteins that are in an interactive population, or that are unique to such population. The PCR products from each of the interactive colonies (the DNA-binding domain-specific amplified products and the activation domain-specific amplified products, respectively) are spotted onto a
20 membrane thus creating an "interactive grid". Preferably, the DNA binding domain-specific products and the DNA activation domain specific products from a single colony are spotted together in a single spot. This interactive grid is then probed with a band of interest that has been identified
25 and isolated through the QEA™ method process. If the band of interest is a band that, through the QEA™ method, has been identified as an interacting band that is present only in one population and not another, this method yields the identity of interacting proteins unique to the population in which
30 such band is present. Probes for this purpose can be prepared by labeling the QEA™ method band(s) of interest with radioisotopes, degoxigenin, biotin (detectable by its ability to bind to streptavidin, e.g., conjugated to an enzyme), fluorescent tags, or other detectable labels known in the
35 art. The spots on the interactive grid are contacted with the probe under conditions conducive to hybridization. Spots that hybridize thus pinpoint the pair of interacting proteins

- 51 -

that are unique to an interactive population (Figure 5). A
sequence analysis of these genes yields the identities of the
interacting proteins.

5   5.2.7.   STATISTICAL CONSIDERATIONS FOR DETECTING
                   ALL POSSIBLE INTERACTIONS AMONG GENES
                   THAT ARE EXPRESSED AT DIFFERENT LEVELS

         In a library of 1×10^6 individual clones, taking
into account that only sense strand cDNAs are cloned and thus
one in every three will be in the proper reading frame, and
10 that each gene has approximately 4 domains, there will be
about 80 copies of each domain of a gene that is expressed at
the high level of 1 in a 1000 transcripts within a cell [(1/3
$\times 1/4 \times 1/1000 \times 10^6$]. After transformation into yeast, if
there are 5×10^5 individual transformants, then there will be
15 40 copies of each domain of a gene that was originally
expressed at a 1 in a 1000 level [80 $\times (5 \times 10^5) \div (1 \times 10^5)$].
These guidelines can be used to calculate the number of
copies of genes expressed at other levels. For instance, it
a gene is expressed at a 1 in 5000 level, a library of 2.5 $\times$
20 10^6 transformants in yeast will be contain roughly
$2.5 \times 10^6 \times (1/3 \times 1/4 \times 1/5000) = 40$ copies of each gene.

         For a given sample size, it is possible to
calculate the number of matings that are expected to yield a
pair of interacting proteins. Suppose that gene X and gene Y
25 are expressed at a level of 1 in 1000, and that domains of
these two genes interact. The fractions of cells bearing the
proper domain of each protein are
Fraction of cells bearing Gene X = 1/(3 $\times$ 4 $\times$ 1000) =
1/12,000;
30 Fraction of cells bearing Gene Y = 1/12,000.
The number of matings that bring together the interacting
domains of gene X and gene Y is
X-Y matings = (total number of matings) x (mating efficiency)
         x (fraction bearing gene X) x (fraction bearing gene Y).
35 Assuming a mating efficiency of 25%, this yields the number
of X-Y matings as:

- 52 -

X-Y matings = (total number of matings) /5.8 x 10^8

Therefore, the total number of matings that must be performed
to expect to see one productive X-Y mating is, on average,
total number of matings = 5.8 x 10^8.

5 This is a statistical estimate of the number of matings;
performing this number of matings will result in a productive
X-Y mating roughly 50% of the time. To raise the probability
of obtaining a productive mating, it is preferable to perform
even more matings. An exemplary goal is a 95% confidence

10 level that an interaction will be retrieved, which requires
3X over-sampling according to probability theory arguments.
For genes expressed at a level of 1 in 1000, the number of
matings for 95% confidence is 1.7 x 10^5.

For genes that are expressed at moderate to low

15 levels, by calculations similar to those described above, the
snumber of matings for 95% confidence is as follows:

Table 1

Expression Level	Number of Matings
1 in 5000	8.5 x 10^5
1 in 10,000	1.7 x 10^{10}
1 in 50,000	8.5 x 10^{10}
1 in 100,000	1.7 x 10^{11}

20

25   Thus, in a preferred embodiment, to detect all
detectable interactions that occur between genes that are
highly expressed in mammalian cells, by assaying interactions
between two populations that are cDNA of substantially total
mRNA from a cell, at least 5.8 x 10^6, or more preferably at

30 least 1 x 10^9, or 1.7 x 10^9 matings between yeast cells in the
preferred yeast interaction mating assays are done. (By way
of clarification, 1.7 x 10^9 matings means mixing 1.7 x 10^9
cells together of each fusion population for a total of
3.4 x 10^9 cells.) The methods described herein allow

35 achievement and selection of these numbers of matings, as
well as the increased number of matings shown in Table 1. In
various specific embodiments, at least 1 x 10^8, 1.7 x 10^9, 8.5

x 10^9, 1.7 x 10^{10}, 8.5 x 10^{10}, or 1.7 x 10^{11} matings are carried
out and Reporter Gene activity is tested for in the mated
cells, per interaction assay.

5   5.2.8. ALTERNATIVE PREFERRED EMBODIMENTS

          This subsection describes specific alternative
embodiments that are generally preferred for the detection
and comparison of protein-protein interactions in the
following circumstance. The embodiments of this subsection
10 are particularly preferred in cases where the binding domain
library has a complexity greater than 10, 1,000, or
1,000,000, and where the number of pairs of interacting
proteins discovered is no more than approximately 10, 50,
100, 200, or 500. However, these embodiments are also
15 applicable to binding and activation domain libraries of
complexities less than 10 and more than 1,000,000 and to less
than 10 or more than 500 discovered interacting protein
pairs. This alternative preferred embodiment is optionally
but preferably associated with certain information-processing
20 steps for recording, comparing, and analyzing the results of
detected interactions. Although applicable in general to the
results of detected protein-protein interactions, these
associated information-processing steps are especially
preferable in cases where one or both libraries have
25 complexity sufficient to result in large numbers of
interactive proteins (i.e., greater than 100, or 200, or
preferably 500 protein-protein interactions), and as will be
apparent to one of skill in the art, these steps are
particularly preferred to record, compare, and analyze the
30 combined results of protein-protein interactions detected
from more than one pair of libraries. Results from multiple
libraries can be from either repetitions of the same pair of
libraries or from different pairs of activation and binding
domain libraries.
35   The current subsection describes generally these
preferred protocol steps to the extent that they differ from
the previously described embodiments. Particular protocols

- 54 -

for performing these steps are presented in the subsections of Section 6.1. Unless otherwise noted, the same choices and alternatives appropriate to the embodiments previously described in Sections 5.1 and 5.2 are also applicable to this

5 embodiment. The following subsection (5.2.9) describes the data-processing aspects of this embodiment. Figure 26 illustrates exemplary orderings of both the preferred protocol steps and the information-processing steps, as well as their interrelation.

10   The steps up to and including the transformation of tne yeast mating strains with plasmid libraries capable of expressing fusion proteins proceed generally as previously described in Sections 5.1 and 5.2. In particular, the previously described choices, namely those of yeast strains

15 with promoter sequences and operably linked reporter genes and of plasmids with marker genes selectable in the yeast strains, are also appropriate to this embodiment. Therefore, by way of example and without limitation, this embodiment is described with respect to a first and a second plasmid

20 library and two yeast mating strains, a and α. When transformed into yeast, the first plasmid library recombinantly expresses *TRP1* and chimeric proteins comprising a *GAL4* DNA binding domain fused to proteins to be assayed for protein-protein interactions, and the second plasmid library

25 recombinantly expresses *LEU2* and chimeric proteins comprising a GAL4 activating domain fused to the same of further proteins to be assayed for protein-protein interactions. The two yeast mating strains are each constructed to be deficient in *TRP1* and *LEU2* and bear reporter genes *URA3*, and/or *HIS3*,

30 and/or *lacZ* whose expression is under control of a *GAL1-10* promoter sequence capable of binding the GAL4 DNA binding domain. This embodiment is adaptable to the other alternatives described in Sections 5.1 and 5.2, in particular to the alternative choices for promoters, reporter genes,

35 selectable marker genes, plasmids, yeast, and so forth therein described.

- 55 -

Where the matrix-mating is performed in
confirmatory step 2606, the plasmids used to construct the
activation and binding domain libraries preferably further
have characteristics which allow them to act as shuttle
5 vectors between the yeast strains used and bacteria such as
E. coli. These characteristics include one or more sequences
permitting replication in bacteria and yeast and one or more
marker genes capable of expression and selection in bacteria
and yeast. The selectable marker genes expressible in
10 bacteria typically express proteins confering resistance to
certain antibiotics.

In more detail, construction of the plasmid fusion
libraries, step 2601 in Figure 26, proceeds as generally
described in Sections 5.1 and 5.2. Genomic DNA or cDNA is
15 prepared from any of various tissues of various organisms
according to appropriate protocols known in the art. For
example, in the case of animal cDNA, mRNA can be extracted
and purified as described in Sections 6.1.3, 6.1.4 and 6.1.5,
and cDNA synthesized as described in Section 6.1.6. The
20 activation and binding domain plasmid fusion libraries can be
constructed according to protocols known in the art. For
example, cDNA, having ends complementary to those produced by
digestion by certain restriction enzymes, such as can be
perhaps produced by ligating short oligonucleotides to
25 previously produced cDNA, can be ligated into plasmid vectors
having appropriate poly-linker sites digested by the same
restriction enzymes. The poly-linker sites are placed in-
frame adjacent to sequences coding for activation or binding
domain protein fragments. For example, the methods of 6.1.6
30 can be used to construct the plasmid libraries.

Transformation of the yeast strains, step 2602 of
Figure 26, also proceeds generally as described in Sections
5.1 and 5.2. Such methods as electroporation,
microinjection, and transformation can be used to introduce
35 the activation and binding domain plasmid libraries into
yeast strains of separate mating types. In an exemplary
method (described in Sections 6.1.2 and 6.1.7), the yeast

- 56 -

strains of separate mating types are transformed with
activation and binding domain plasmid libraries by lithium
acetate treatment followed by heat-shock.

Following transformation step 2602 is negative
5 selection step 2603. This step screens out those yeast
transformants bearing binding domain plasmids in which the
reporter genes are fortuitously activated by the fusion
protein bearing the binding domain alone. Such fortuitously
activating transformants can make impractical the task of
10 finding a tiny number of colonies truly positive for protein-
protein interactions among an overwhelmingly large number of
falsely positive colonies produced from libraries of large
complexity. For example, each such fortuitously activating
binding domain transformant will mate with any activation
15 domain transformant to form falsely positive progeny which
will grown on a medium selective for reporter gene
activation. Therefore, the greater the complexity of the
activation domain library, the more such false positive
progeny will be formed from each such fortuitously activating
20 binding domain transformant. Additionally, fortuitous
activation can occur at a rate up to 1-5% among all binding
domain transformants. Therefore, the greater the complexity
of the binding domain library, the more such false positive
progeny will be formed. For binding domain libraries with
25 complexities of greater than 10^5, 10^6, 10^7, or even 10^8, it is
preferable that the rate of fortuitous activation be below at
least 10^{-5}, more preferably less than approximately 5×10^{-6},
and most preferably less than approximately 1×10^{-6}. The
"rate of fortuitous activation" means the fraction of binding
30 domain fusion transformants that activates reporter genes in
the absence of any protein-protein interaction.

A negative selection protocol preferred for use
with this embodiment achieves a much reduced fortuitous
activation rate by combining separate and independent
35 negative selection steps. It is important that such separate
negative selection steps be independent in order that their
negative selection effects be cumulative. The preferred

negative selection protocol achieves a fortuitous activation
rate of preferably less than approximately 5×10^{-6}, or less
than approximately 4×10^{-6}, or less than approximately 3×10^{-6}, or less than approximately 2×10^{-6}, or more preferably

5   less than approximately 1×10^{-6}, or even less. In a
preferred embodiment, where *URA3* is a reporter gene two or
more passages are made on media containing 5-fluoroorotic
acid (5-FOA) (the chemical agent creating the toxic
environment for *URA3*), which inhibits or kills URA+ cells.

10   In a first passage, binding domain transformants are plated
on media selective for the binding domain plasmid and
containing 5-FOA. After a sufficient time for growth,
resulting colonies are replica plated onto similar selective
media also containing 5-FOA. It has been found that two

15   passages by replica plating achieve a fortuitous activation
rate of no more than approximately 1×10^{-6}. Further passage
via replica plating are possible, and can be performed if
fortuitous activation rate greater than the preferred rate is
found.

20   Replica plating is a preferred embodiment of the
general method of achieving independent negative selection
steps according to this invention. The general method
proceeds by using any appropriate means to definitively
separate those cells, which are actively growing in a toxic

25   environment, from substantively all other cells, including
dead cells, cells which are living but not viable, and
importantly, cells which are dormant in the toxic environment
but still viable and capable of future growth in a non-toxic
environment. By way of example, it has been found that an

30   important, although small, fraction of yeast cells in a toxic
environment, such as a medium containing 5-FOA for URA+
cells, are not killed, but merely become dormant yet viable.
Such viable dormant cells are fully capable of resuming
normal growth upon being rescued to a new non-toxic

35   environment. In particular, in the case of an organism, such
as yeast, for which cells growing on a plate create colonies
forming a heap above the surface of the medium, actively

- 58 -

growing cells on a plate containing a medium with a toxic
chemical agent create such heaped-up colonies, while dormant
cells remain on the surface of the medium. Accordingly,
definitive separation of actively growing cells can be

5 achieved by physically removing cells from the heaped-up
colonies, and preferably from the tops of these colonies,
without removing cells from the surface of the medium.
Careful replica plating, the preferred means, reliably and
economically removes only cells from the tops of heaped-up

10 colonies. Alternatively, other physical means can be used to
remove cells from heaped colonies, such as careful colony
picking, perhaps by a laboratory robot. On the other hand,
scraping cell from the surface of such a medium removes both
growing cells and dormant cells, and therefore, is

15 ineffective in achieving independent negative selection
steps. The dormant cells later resume growth in a non-toxic
environment. Also, growth in successive liquid media having
the toxic agent, without additional plating, does not achieve
independent selection and improved negative selection rates.

20   After careful separation of actively growing cells,
their further growth in a further toxic environment results
in further and independent selection by killing remaining
sensitive cells. Dormant cells which escaped death in the
previous toxic environment will not again escape selection in

25 this further toxic environment, since substantially none of
these cells are transferred to the second toxic environment.
Accordingly, the results of both selection steps combine to
result in a much reduced fortuitous activation rate.

          Alternatively, other reporter genes and associated

30 toxic environments, as described in Section 5.1 or known in
the art can be used in this protocol. It is preferable that
all such combinations achieve a rate of fortuitous activation
of less than 5×10^{-6} and more preferably less than
approximately 1×10^{-6}. For example, an alternative protocol

35 can use two or more passages by replica plating in the
presence of cycloheximide where CYH2 is present as a reporter
gene (cycloheximide is the toxic chemical agent for *CYH2*) in

the yeast.   Alternatively, the two passages can involve
growth on media having different compounds that are toxic
upon fortuitous activation of separate reporter genes.   For
example, where both *URA3* and *CYH2* are used as reporter genes,
5 a first passage can be on media containing 5-FOA and a second
passage can be on media containing cycloheximide.   In a
further alternative, where two reporter genes having
difference toxic environments are used, one or more passages
can be on media with both toxic environments.   For example,
10 where both *URA3* and *CYH2* are used as reporter genes, one or
more passages can be on a medium containing both 5-FOA and
cycloheximide.   In all of these alternatives, as described,
only actively growing cells must be carefully selected for a
further negative selection step.

15   A further negative selection step, called bait
validation, is preferred in the case of libraries of limited
complexity.   Such libraries have a complexity preferably less
than approximately 500, or less than approximately 200, or
less than approximately 100, or most preferably less than
20 approximately 50.   The goal of the step, in the case of
binding domain libraries, is to provide a further screen for
fortuitously activating binding domain fusion proteins, and
in the case of both binding domain and activation domain
fusion proteins, is to provide a screen for "sticky" fusion
25 proteins (see, also, Section 6.1.13.2).   Although a
particular fusion protein may activate reporter genes due to
true protein-protein association, this association may be
non-specific.   Since such non-specific association may be of
less interest than specific association between proteins, it
30 may be advantageous to remove library members expressing such
sticky fusion proteins before a full mating.   After a full
mating and positive colony selection, the matrix-mating
protocol described subsequently performs a similar screen for
fusion proteins that associate non-specifically with many
35 other partners in a particular mating.

          For the bait validation protocol, fortuitously
activating binding domain fusion proteins and sticky fusion

proteins are recognized by the rate of reporter gene
activation during a test mating (as described below). As
used herein, the rate of reporter gene activation in a mating
is the fraction of diploid cells in which one or more
5 reporter genes are activated. Fortuitously activating
binding domain fusion proteins are recognized by a rate of
reporter gene activation that is close to 1, e.g., greater
than or approximately 0.5. Sticky fusion proteins are
recognized by a rate of reporter gene activation that is
10 anomalously high compared with the expected rate, as
determined by observations of similar matings. For example,
in matings of mammalian and, particularly, of human samples,
it has been observed that the protein-protein association and
reporter gene activation is typically less than approximately
15 10^{-6} (i.e., reporter genes are activated about 1 diploid cell
in 1,000,000 diploid cells). Accordingly, for similar
matings, a sticky fusion protein is indicated by a rate of
reporter gene activation preferably greater than
approximately 10^{-5}, or preferably greater than approximately
20 10^{-4}, or more preferably greater than approximately 10^{-3}.
Since it is generally advantageous to detect as many weak
protein-protein interactions as possible, a library member
with a rate of reporter gene activation in a test mating of
greater than a threshold of approximately 10^{-3} is considered
25 "sticky." Where only stronger protein-protein interactions
are of interest, fusion proteins with activation rates
between 10^{-3} and 10^{-4} (or 10^{-5}) can also be considered "sticky."
Limited-complexity-library members are considered validated
for performing full library mating only if they are neither
30 fortuitous activators nor are sticky, that is if their
reported gene activation rates are less than the appropriate
thresholds.

An exemplary protocol for bait validation performs
a separate mating, according to the protocols described
35 herein, of each member of the limited complexity library with
a sample of the more complex library. For example, each
member is mated preferably with between approximately 10,000

and 100,000 colonies from the more complex library, and most
preferably with approximately 50,000 colonies. The
approximate rate of diploid colonies which are also positive
for reporter gene activation for the member is observed.
5 Only library members which meet the preferred rate of
reporter gene activation (where weak protein-protein
interactions are of interest, a rate of greater tan 10^{-3}) are
selected for full mating.

   This invention also comprises other negative
10 selection techniques performed before a full mating, directed
to removing from the full mating any fusion proteins that
fortuitously activate reporter gene expression and/or have
non-specific (sticky) association with other proteins, that
will be apparent to those of skill in the art upon reviewing
15 this disclosure.

   Following transformation and negative selection,
the libraries of yeast transformants are mated and colonies
selected for activation of the reporter genes in step 2604 of
figure 26. In general, a mating protocol useful in these
20 embodiments has the following preferable characteristics.
First, it is preferable that the large numbers of cells
necessary for good mating of complex libraries, according to
the statistical estimates of Section 5.2.7, can be mated
using only a limited number of plates, and limited media and
25 mating resources. Second, mating conditions chosen promote
cell mating but inhibit cell doubling. Accordingly, each
separate mating event constituting a protein-protein
interaction is more likely to produce only a single resulting
colony upon selection. Third, also for good statistical
30 sampling, the mating efficiency, the percentage of diploids
formed, is high.

   Generally, high mating efficiencies are produced
when well mixed yeast cells of the two mating strains are
maintained in fixed and close contact, as occurs when the
35 mating cells are packed together and retained on various
solid supports. Accordingly, mating on the surface of plates
or filter discs is preferred, with filter discs being more

preferred due to their ability to pack together and to mate a greater number of cells per disc.

One aspect of this invention is the discovery that the disclosed filter-disc mating protocol permits
5 significantly higher cell densities during mating than can be achieved with prior mating protocols, in particular by mating on the surface of a plate. In particular, filter-disc mating can achieve approximately at least 5×10^4, at least 1×10^5, at least 1.5×10^5, preferably 3.5×10^5, and up to $4-6 \times 10^5$
10 cells per square millimeter on the filter-disc during mating. Mating cell densities above $4-6 \times 10^6$ are less advantageous since mating efficiency declines. These densities correspond to at least approximately 3×10^8 cells, to at least approximately 6×10^9 cells, to approximately 1×10^9 cells,
15 to approximately 2×10^9 cells, and up to approximately 3.5×10^9 cells per 90 mm filter disc, respectively (obtained by multiplying the cell densities by the approximately 6400 square millimeters in a 90 mm filter disc). According to the preferred protocol, cells can be packed to these densities on
20 a filter-disc by vacuum-assisted filtration from a culture of known cell density by using various standard filtration apparatuses. Filter discs of different diameters can accommodate appropriately scaled numbers of cells. Prior methods can typically accommodate, at most, a mating cell
25 density of 6×10^3 cells per square millimeter (for example 1×10^8 cells on a 150 millimeter plate).

Cell doublings during the mating in a filter disc are limited by maintaining the mating cells in an environment of a rich but dilute medium, as can be readily achieved by
30 placing filter discs with the packed yeast cells cell-free side down on the surface of a plate with rich medium (e.g., the YPAD medium described in Section 6.1, supra). Mating efficiently is also promoted by "boosting" the cells with a short growth period on rich medium prior to mixing and
35 mating. In contrast, plate mating places the cells on a rich medium resulting, typically, in several cell doublings and several colonies for each positive mating event.

- 63 -

As in the protocols of Sections 5.1 and 5.2, mated
cells are harvested and further plated out on media selective
both for activation and binding domain plasmids, and thus for
diploid yeast cells, and for activation of the reporter
5 genes. Cell from positive colonies are taken by, e.g.,
picking from the plates containing medium selective for the
presence of both plasmids and reporter gene activity (mating
plates) and stored in individual cultures selective for both
plasmids, which are, for example, arrayed in 96-well plates,
10 384-well plates, or other convenient storage format. Cells
for further analysis of the positive colonies can be removed
from the storage cultures. It is advantageous for removing
colonies from the mating plates that the number of expected
positive colonies as well as the total number of diploid
15 cells per plate be controlled. Too many colonies per plate
makes difficult picking colonies from mating plates to place
them in storage cultures. Too few colonies per plate wastes
mating plates. In a particular embodiment, directed to
automatic colony picking by robot apparatus guided by an
20 automatic vision system, a preferred number of colonies per
plate is approximately 50-100 and a preferred number of
diploids per plate is less than approximately 10^3.

These plating targets are attained by estimating
the expected percentage of diploid cells among all the mated
25 cells and by estimating the expected rate of protein-protein
interactions among all the diploid cells. One of skill in
the art knows how to plate appropriate dilutions of the
harvested, mated cells in view of these fractions and of a
measured cell density. The percentage of diploids, or the
30 mating efficiency, can be estimated by plating serial
dilutions of the mated cells onto plates selective for each
of the plasmids and for both of the plasmids (for example,
according to the protocol in Section 6.1.1). The expected
rate of protein-protein interactions can be estimated from
35 experience with similar libraries. In the case of libraries
derived from total mRNA of human cells, the rate is often
approximately 10^{-7}, or at least between 10^{-6} and 10^{-8}.

The positive colonies harvested at the end of step
2604 can be processed according to subsequent steps and
protocols, including the characterization of the fusion
protein produced at step 2605, confirmatory tests at step
5 2606, and other further steps indicated at 2607. The
confirmatory tests screen for false positive colonies due to
fortuitously-activating binding domain plasmids (plasmid
drop-out test) and eliminate non-specifically interacting
proteins (matrix-mating test). The other further steps are
10 described in Sections 5.1, 5.2, and 5.3 and illustrated in
Figures 1, 3, 5, and 6, and include screening for inhibitors
of protein-protein interactions (described in Section 5.4),
finding lead compounds for drugs that inhibit protein-protein
interactions, finding stage or tissue specific protein-
15 protein interactions, and so forth.

        These subsequent steps can be performed in any
order or even eliminated if not needed. The order shown in
Figure 26 is the preferred order, especially where associated
information processing steps assist the analysis of
20 interesting interactions. In the preferred order, fusion
protein characterization is performed first and produces
input that the information processing steps use to control
performance of the confirmatory steps, which are performed
second. Other orderings can include performing all these
25 steps in parallel, performing confirmatory tests in advance
of fusion protein characterization, eliminating the further
steps, or other variations.

        Step 2605 characterizes the fusion proteins in each
of the positive colonies harvested from the mating step.
30 Information produced in this step is input, as represented by
input arrow 2608, to the information processing steps which
generally act to further characterize the interaction.
Sections 5.1 and 5.2 describe several methods for this
characterization. The pooling and deconvolution described
35 therein are preferably not applied to this embodiment. Since
it is anticipated that less than approximately 10, or 50, or
100, or 200, or 500 positive colonies are found, the

- 65 -

identification efficiencies provided by pooling and deconvolution are less advantageous to this embodiment.

According to Sections 5.1 and 5.2, and also in this embodiment, analysis of separate and individual colonies
5 proceeds, preferably, with a first amplification step followed by a second characterization step. The amplification step specifically amplifies the variable inserts coding for the interacting protein fragment in the fusion proteins, by, in the case of PCR amplification, using
10 primers designed to hybridize to regions flanking the variable inserts. The second step, which characterizes the amplified inserts, can be by direct sequencing, or by QEA or SEQ-QEA methods (described in Section 5.4), or by other methods. Direct sequencing is preferred in this embodiment,
15 especially where adequate sequencing facilities are available, and the sequence data is directly input to the information processing steps. Direct sequencing can be by any method known in the art, but is preferably according to the Sanger chain-termination method using ddNTPs labeled with
20 four distinguishable dyes and followed by electrophoretic separation of the sequencing fragments. If QEA or SEQ-QEA methods are employed, the QEA signals (described in Section 5.4) produced are input to the information processing steps, and gene identification is preceded by the gene finding
25 methods described in Sections 5.4.5 and 5.4.6.

In detail, the first PCR amplification step preferably uses DNA templates produced from yeast obtained from the positive colony storage. The DNA templates are freed of cellular debris by extracting DNA from the results
30 of cell lysis and proteolysis (as described in Section 6.1.8). Preferred hot-start PCR protocols are also described in Section 6.1.8. A most preferred protocol separates components of the PCR reaction mix by a solid wax layer, so that no amplification can occur until the wax layer is
35 melted. To start amplification, the PCR reaction mix components are pre-heated, the wax layer is melted, and thereby, the amplification is hot-started. This latter

- 66 -

protocol is easily adapted to performance by standard
laboratory robots.

Finally, step 2606 confirms certain aspects of
positive colonies found after the mating step.   In
5 particular, the plasmid drop-out test performs a protocol
(described in detail in Section 6.1.13.1) that separates
false positive colonies, due to reporter gene activation
solely by the binding domain fusion protein, from true
positive colonies, in which reporter gene activation requires
10 protein-protein association. In embodiments accompanied by
information processing, performance of this steps is
controlled, as indicated by control arrows 2610 and 2612, by
assessment of the quality and biological significance of a
particular interaction at step 2618 or by browsing the
15 database of interactions at step 2620. Results of these
confirmatory steps are input, according to input arrow 2611,
to the information processing.

Briefly, the plasmid drop-out protocol grows cells
from a positive colony, first, in rich complete medium, and
20 second, in medium selective for the binding domain plasmid in
order to select for drop-out of the activation domain
plasmid. The selected progeny are tested for such drop-out
by lack of growth in a medium selective for the activation
domain plasmid. Progeny cells lacking the activation domain
25 plasmid are then assayed for activation of one or more of the
reporter genes. Any positive colonies having reporter genes
activated only by the binding domain plasmid are considered
false positive for protein-protein interactions.

The matrix mating test performs a protocol
30 (described in detail in Section 6.1.13.2) that assays for the
specificity of observed protein-protein interactions.
Generally, this test reconstitutes a second two-hybrid
interaction test using only the activation and binding domain
plasmids from colonies positive during a first interaction
35 test. If a protein-protein interaction is specific, then it
is expected that the activation and binding domain plasmids
bearing the components of the specific interaction will form

- 67 -

a positive colony only when they are mated together, and will
not form positive colonies when they are mated with other
plasmids. On the other hand, if a protein component
interacts non-specifically, then it is expected that the
5 plasmid bearing that component will form positive colonies
with many other plasmids. The interaction test is
reconstituted, in summary, by rescuing and maintaining
plasmids from the positive colonies into a bacteria, such as
E. coli. Accordingly, it is advantageous that the plasmids
10 used have characteristics of shuttle plasmids. Separate
yeast mating strains are transformed with the activation and
binding domain plasmid DNA extracted from the bacteria. The
strains are mated and grown on media selective for the
reporter genes. In a particular embodiment, yeast cells
15 containing the different plasmids are grown on lines arranged
in a grid that intersects (a matrix). A positive protein-
protein interaction appears as growth as the intersection of
the two lines having the plasmids bearing the components of
the interaction.

20   This invention also comprises other negative
selection techniques performed after a full mating, directed
to removing from the selected positive colonies any colonies
with fusion proteins that fortuitously activate reporter gene
expression and/or have non-specific (sticky) association with
25 other proteins, that will be apparent to those of skill in
the art upon reviewing this disclosure.

        This embodiment further comprises observation of
"bi-directional" interactions (also called herein "bi-
directional screens"). Two fusion inserts, a first and a
30 second insert, participate in a bi-directional interaction if
they are observed to interact under the following two
conditions or directions: one, with the first insert in a
binding domain fusion protein library and the second insert
in an activation domain fusion protein library in a first
35 direction; and two, with the first insert in an activation
domain fusion protein library and the second insert in a
binding domain fusion protein library in a second direction.

Bi-directional interactions can be discovered by performing an interaction detection assay twice, first with a pair of libraries constructed to have the inserts in either the first or the second direction, and second, with another pair of

5 libraries constructed to have the inserts in the other direction. Finding two fusion inserts in a bi-directional interaction increases the likelihood that the observed interactions is experimentally significant, and not an artifact of the fusion libraries.

10   In summary, a particular embodiment of this preferred alternative embodiment of this subsection proceeds according to the following steps: construction of fusion plasmid libraries; transformation of yeast strains; negative selection of the binding domain library; mating of the yeast

15 strains; selection of colonies positive for activation of the reporter genes; characterization of fusion protein from positive colonies; confirmatory tests such as plasmid drop-out and/or matrix-mating; and optional further steps. Where, as is preferred, information processing accompanies these

20 steps, the fusion protein characterization and the confirmatory steps input information into information processing functions for further control of these same steps and for recording, analysis, and comparison of protein-protein interactions observed.

25

5.2.9. INFORMATION PROCESSING ASPECTS OF DETECTING PROTEIN-PROTEIN INTERACTIONS

The information-processing aspects of detecting protein-protein interactions record, compare, and analyze

30 protein-protein interactions detected in experiments (also referred to herein as "screens" or "matings") involving one or more pairs of libraries. These information-processing aspects are important to manage the large amounts of information generated from interactions detected in complex

35 libraries, and especially from interactions detected in many pairs of complex libraries. Although the information-processing aspects are described primarily with respect to

the alternative preferred embodiment of section 5.2.8, they
are applicable to all embodiments of identification and
comparison of protein-protein interactions according to this
invention. Further it will become apparent to those of skill
5 in the art, that the data structures and processes described
are also usefully applicable to other biological systems (and
to non-biological systems) consisting of many pair-wise
interacting components. They are even more applicable to
such of those systems where the pair-wise interactions are
10 determined by components which can be systematically sampled
according to geometrically comparable parameters, such as
linearly arrangeable nucleotide or amino acid sequences.

In this subsection, the information-processing
aspects are described, first, with respect to their functions
15 and relevant data classes, and second, with respect to
detailed structures of their databases, detailed sequences of
information-processing steps, and their relation to
accompanying protein-protein interaction detection.

The information-processing aspects provide, among
20 others, three groups of functions and employ, among others,
three classes of data. The first group of functions is
directed to identifying, if possible, the genes coding for
the protein fragments which have been found to interact, or,
at least, produce colonies positive for reporter gene
25 activation. This group also includes functions for
organization and storage of data returned from the
experimental protocols for detecting protein-protein
interaction, for example, the data describing interaction
experiments performed and results of fusion protein
30 characterization from positive colonies. The second group of
functions is directed to quality control of the results of
protein-protein interaction detection. It assists a user to
assess the biological meaning of each positive colony, for
example, candidate identifications of the genes coding for
35 the interacting fusion fragments found, and to identify the
biological context of the interactions detected. These
functions also assist with management of steps of the

experimental protocols, in particular, selection of the confirmatory tests to be performed in view of the biological significance and context found for an interaction. Such management is generally called "workflow." The third group

5 of functions assembles interactions deemed significant, for example, because they are detected from two or more separate library mating experiments, and provides facilities for review and analysis of the assembled protein-protein interactions. In particular, this group also provides for

10 assembling detected interactions between pairs of proteins into pathways linking multiple proteins and for discovering the domains in the proteins responsible for observed interactions.

With regard to the classes of data employed, the

15 first class includes principally raw data describing and/or returned from each protein-protein interaction experiment. The data describing a particular experiment includes at least unique identifiers for each mating experiment and for each colony found to oe positive for reporter gene expression.

20 This data optionally also describes the DNA libraries used to construct the plasmid fusion libraries and the precise materials, methods, and conditions used in this mating. Data returned from a particular experiment includes at least sequences of the fusion inserts (the library DNA sequences

25 joined with the activation domain and binding domain sequences in the plasmid libraries) found in positive colonies, or in the case of QEA analysis, the QEA signals generated from the amplified fusion fragments.

The second data class supplements the first class

30 by adding both organization and indexing components built over the first class of data, in order to make it accessible for easy reference, and also candidate identifications of the genes coding for the positive fusion inserts. If no currently known gene codes for a particular fusion insert, an

35 internal accession number is generated to refer to the putative new gene and the closest homologous genes are recorded.

Finally, the third class of data records all the
distinct protein-protein interactions found, each of which is
characterized at least by the genes coding for the particular
interacting proteins. For each such protein-protein

5 interaction (referred to herein as an "interaction"), this
class also includes data describing all the individual
positive colonies (referred to herein as "interactants")
whose two fusion inserts (referred to herein as an
"interacting pair") are fragments of the proteins coded for

10 by the genes characterizing the parent interaction. This
third class is particularly useful and can be further
processed, as described subsequently, to yield useful
additional information.

Prior to describing the processing steps in more

15 detail, a preferred and exemplary hardware and software
implementation of the functions and data classes is
presented. It is understood that this invention includes
other hardware and software implementations that achieve
equivalent functions. The individual groups of functions and

20 certain components of these groups are preferably implemented
as independent programs which are coordinated by client-
server style communication. Such client-server
implementations are known in the information-processing arts.
The individual client and server components are distributed

25 on hardware platforms in a convenient and economical manner.

Figure 27 illustrates an exemplary hardware system
configuration implementing for an exemplary distribution of
client-server function. Computer 2702, which can be two or
more computers, hosts programs implementing the previously-

30 described groups of functions and connects databases and
files storing the classes of data. As is generally
understood in the art, information relating to the entities
in the files and databases of this invention is represented
and stored in digital form. The digital representation can

35 be according to any convenient code known in the art. The
first class of data is typically stored largely in structured
user-maintained files 2708, for example in descriptive text

- 72 -

files. Preferably, the second and third classes of data are largely stored in relational databases. Identification database 2706 stores the second class of data, and interaction database 2707 stores the third data class. A
5 preferred relational database system (version 7.0 or preferably 7.3) is available from the Oracle Corporation.

Computer 2701 connects to sequence database 2705 which is consulted in the process of determining candidate gene identification. Where fusion inserts are sequenced, as
10 is preferred, computer 2701 searches for database sequences homologous to insert sequences. Where QEA signals are available, computer 2701 performs a database search process similar or equivalent to that described in Section 5.4.5 (see especially Section 5.4.5.1).

15 User computer 2703 connects to user display and keyboard 2709 in order to provide user access to the information processing aspects of this invention. Typically, multiple users access the information-processing system from multiple user computers similar to computer 2703. Where
20 information-processing functions include workflow management components that control steps of the interaction experiments, user computers can be made available to the laboratory technicians responsible for actually performing the protocol steps. Where the steps involve routine manipulations,
25 laboratory robots 2710 can be directly interfaced to the user computers. Such robots can be controlled by and can return data to the information processing functions. For example, positive colony identification and picking can be performed by a robot.

30 The computers are connected by communication links 2704, which are adapted to the actual physical distribution of the computers as is common in the art. When the computers are collocated, link 2704 can be a local area network; when the computers are remotely located, link 2704 can be, for
35 example, the Internet. Combinations of networks can be used when computers are variously located.

In detail, system computers are appropriately sized
according to their processing loads, but are preferably at
least 166 Mhz or greater Pentium™-based computers (or
computers of equivalent performance based in Sparc™ or Alpha™
5 processors). The system computers are provided with standard
software components, including an operating system, which can
be a version of UNIX (for example, one of the versions
available from Sun Microsystems) or one of the Windows™
family of operating systems from the Microsoft Corporation
10 (Windows NT™, or Windows 95™). Implementation languages can
be general purpose languages, such as C, C++, Java™, language
directed to relational database manipulation, such as PL/SQL™
(Oracle Corporation), or similar language. The preferred
language for graphical presentation aspects of these methods
15 is Java™, and the preferred language for relational database
manipulation and text screen formatting is PL/SQL™.
Presentation services at the user computer are preferably
provided by an internet browser, such as NetScape™ from the
NetScape Corporation, or other equivalent program capable of
20 interpreting HTML formatted screens.

      This invention also includes computer readable
media which contain computer-readable instructions capable of
causing one or more computers to perform the processes of
this invention. Such media include magnetic discs and tapes,
25 optical discs, and other media types. The computer-readable
instructions on these media include both instructions for
performing the processing steps of this invention and also
instructions for defining and establishing the files and
databases of this invention.
30

      5.2.9.1. <u>IDENTIFICATION DATABASE AND PROCESSING</u>
      In this and the following subsections, the
information-processing functions and data classes are
described in more detail. First, the identification database
35 and its processing functions are described. Next, the
interaction database and its creation and update are
described. Lastly, functions are described which are capable

of deriving further information beyond that literally contained in the interaction database. Generally, the right hand column of Figure 26 illustrates an implementation of these information-processing steps and their relationship to

5 the steps for protein-protein interaction experiments.

Identification database 2617 for a protein-protein interaction experiment is created by gene identification step 2616 using input data 2614 and (external) sequence databases 2615. Input data 2614 describes the protein-protein

10 interaction (mating) experiment and characterizes the fusion protein inserts from colonies positive for reporter gene activation. A description for a protein-protein interaction experiment includes at least a unique identifier that permits efficient retrieval of all information relating to this

15 experiment. Further descriptive information includes, most preferably, information on the DNA source libraries from which the activation domain and binding domain plasmid fusion libraries were made. DNA library description can recite animal and tissue origin, library complexity, disease state

20 and/or treatment information, if any, methods of library production, storage location of library samples, and so forth. Additional experimental description information can include the precise and particular materials, methods, and conditions used in the protein-protein interaction protocols.

25 This descriptive information can be stored in coded or in free-text form, in files or in a database system, and can be advantageously indexed according to certain fields for rapid retrieval. For example, all data relating to a particular mating experiment is easily retrievable by using the unique

30 experimental identifier. It is also advantageous that data from all experiments relating to selected libraries, species, tissue types, diseases, treatments, and so forth be similarly easily retrievable by searches on the corresponding fields.

In addition, input information 2614 includes data

35 from each colony found to be positive for reporter gene activation. Each positive colony is assigned a unique identifier, and information obtained from that colony is

indexed for rapid retrieval using this identifier. The
combination of mating experiment identifier and colony
identifier for a positive colony is preferably unique among
all the mating experiments and positive colonies data stored
5 in a particular implementation of these information-
processing aspects. Data available for a positive colony
characterizes the fusion inserts found in the colony, and
preferably, also includes management information such as the
physical storage location of the colony and so forth. The
10 physical location for a colony indicates to a laboratory
technician the location from which to retrieve cell samples
for further experimental steps.

Preferably, nucleotide sequences characterize the
fusion inserts found in a positive colony. Such sequence
15 data is commonly provided by commercially-available
sequencing machines in various output formats. Most simply,
the sequences of the activation domain and binding domain
fusion inserts can be simply stored as, e.g., a string of
nucleotide identifiers along with an indication of the
20 correct reading frame. Where the QEA or the SEQ-QEA methods
are used, the fusion inserts are characterized by QEA
signals. QEA signals, described in detail in Section 5.4,
comprise three pieces of information, namely, the sequences
of two subsequences present in the fusion insert (each having
25 a length of, typically, 4 to 6 nucleotides) and the distance
between these subsequences. In the case of SEQ-QEA signal,
the subsequences are typically from 8 to 12 nucleotides long.
All data for a particular colony is preferably easily
retrieved using its colony identifier.

30   Identification step 2616, which creates an
identification database for a particular mating experiment,
also refers to certain external databases, primarily external
sequence databases 2615. Representative external sequence
databases are available from governmental organizations (for
35 example, GenBank from the National Institutes of Health and
similar databases available from the European Molecular
Biology Laboratory) and from private organizations. By way

of example, without limitation, the following description in this subsection is in terms of GenBank available at the Internet address: "http://www.ncbi.nlm.nih.gov"

Prior to describing the processing which creates
5 identification database 2617, information present in this database is described. Exemplary contents of an identification database are presented in the following Table 1A.

10

15

20

25

30

35

TABLE 1A: IDENTIFICATION DATABASE

FIELD	DESCRIPTION
mating experiment identification	Appropriate unique identification of the mating experiment with links to description of libraries used, precise protocols, and so forth
positive colony identification	Appropriate unique identification of this positive colony (in particular for future retrieval from storage)
A list of candidate homologues for activation domain fusion protein insert	
gene-AD	Identity of homologue for activation domain fusion insert (database or internal accession number)
description of gene-AD	Name, species origin, tissue origin, and so forth
3'-5' position-AD	Location of fusion insert sequence on the homologue sequence (nucleotide positions of fragments ends)
score-AD	Probability of homologue (e.g., BLAST probability)
A list of candidate homologues for binding domain fusion protein insert	
gene-BD	Identity of homologue for binding domain fusion insert (database or internal accession number)
description of gene-BD	Name, species origin, tissue origin, and so forth
3'-5' position-BD	Location of fusion insert sequence on the homologue sequence (nucleotide positions of fragments ends)
score-BD	Probability of homologue (e.g., BLAST probability)

The mating experiment and colony identification fields contain their previously described identifiers. For each positive colony, this database includes lists of one or more candidate genes that have been determined to possibly code for the inserts in the activation domain and binding domain fusion proteins. "Genes" are used herein to refer to nucleic

acid coding sequences, which can be, for example, cDNA or genomic. These genes are identified by their database "accession numbers," for example the widely-used GenBank accession number. It is well known in the art that accession

5 numbers can redundantly identify certain sequences, for, at least, the reason that sequences of various fragments of the same naturally-occurring nucleic acid can have been entered multiple times. To obtain a unique gene identifier, it is preferred to use the accession number of the full length

10 coding sequence for the gene, or at least, the accession number for the longest fragment including the section found to be a candidate for the fusion insert. For each candidate gene, the identification database preferably includes at least certain descriptive information, the location of the

15 insert in the putative gene, and a score of the degree of homology with the insert. Descriptive information, in turn, preferably includes species origin and, optionally, an indication of the general function, if known, of the protein coded by the gene, for example, a cell cycle protein, a

20 signaling pathway protein, transcription factor, an enzyme, or structural protein and so forth. The location of the insert on the gene sequence is described by providing, e.g., the nucleotide numbers in the gene sequence of the 3' and 5' ends of the insert. The score field is an estimate of the

25 likelihood that this gene actually includes the insert sequence by providing, e.g., a degree of sequence homology or a probability that the two sequences are randomly related.

In alternative embodiments, the identification database can also include such other information that will

30 aid in quality control step 2617, as will be apparent to those of skill in the art. For example, from protein databases additional information about proteins known to be associated with the gene can be added. Information can also be added from still other databases such as, e.g., literature

35 databases searched with gene name or accession number.

The identification database is preferably stored in a relational format in an appropriate normal form, as is

common in the art.   Tables can be defined for relating
experiments and their positive colonies and for relating
colonies and their candidate genes.   Alternatively, this
database can be stored in other database formats, or in a set
5 of user maintained files.   In a further less preferable
embodiment, the data content of the identification database
can be stored as files in a raw or unprocessed form, perhaps
with distinctive filenames.

Gene identification step 2616 creates or updates
10 the identification database for a particular mating
experiment.   According to the above exemplary table
definitions, first, tables relating all the positive colonies
to the particular mating experiment are loaded.   Thereby, the
unique colony identifiers, and perhaps colony descriptive
15 information, are related to the unique experiment identifier.
Next, tables relating the insert sequences in the positive
colonies to their candidate genes are loaded.   During this
step, the candidate sequences need to be determined.

The determination of candidate genes proceeds, in
20 the preferred embodiment, by using one of the several
homology search programs existing in the art, and in the
alternative, by using the QEA experimental analysis methods
described in Section 5.4.5.1.   In the preferred embodiment,
candidate genes are selected by searching a sequence database
25 with a homology search program using the determined fusion
insert sequences as queries.   These programs often function
in a client-server mode, accepting formatted query sequence
queries, referencing a nucleotide sequence database, and
returning output text files describing the results of the
30 homology search.   The output text files typically contain a
list of homologous sequences (genes) from the sequence
database together with, for each sequence, an indication of
the likelihood of the homology and an indication of how the
query maps onto the sequence.   A preferred homology program
35 is BLAST (Altschul et al., 1990, Basic Local Alignment Search
Tool, _J. Mol. Biol._ 215:403-410) which is available at the
Internet address "http://www.ncbi.nlm.nih.gov."   BLAST

returns text files with the preferred information, that is
sequence accession number, query sequence location, and a
homology score (multiple possible locations and associated
homology scores can be provided). A copy of BLAST along with
5 a sequence database can be loaded onto local computers.

In this embodiment, when the fusion-insert sequence
data becomes available for the positive colonies of an
experiment, it is collected (for example by retrieving output
text files from BLAST located using the experiment
10 identifier) into a set of queries formatted for the BLAST
program, one query for each of the fusion insert sequence.
The queries are sent to an instance of BLAST and the output
text files are received and stored. These output text files
are then parsed in manners well known in the arts (for
15 example by a program in the PERL language), the relevant data
extracted, and the identification database accordingly
updated. Alternatively, the output files can be used in the
received format, perhaps indexed by colony identifier for
easy retrieval.

20   In the case of fusion insert characterization by
QEA signals, gene identification proceeds according to the
following steps. A sequence database is searched using the
QEA (or SEQ-QEA) signals as queries according to the
processes described in Section 5.4.5.1. The output is a set
25 of candidate sequences (genes) that include fragments
generating the same signals as generated by the fusion
inserts. For each candidate sequence, an approximate
position of the fusion insert, even though not sequenced, can
be found from the positions of all the fragments of the
30 candidate gene known to generate the observed signals.
Figures 17A-F and the accompanying description illustrate how
the observed signals correspond to fragment with particular
positions on the candidate sequence. Since the signals
generated by the fusion insert originate from fragments at
35 known locations on the candidate sequence, the fusion insert
must include at least the overlap of all the fragments.
Thereby, overlapping on each candidate gene all the fragments

- 81 -

corresponding to signals generated from a fusion insert leads
to an approximate position of this insert on the candidate
gene. In this embodiment, the only accessible homology
information of the candidate sequence is according to the
5 methods of Section 5.4.5.3, that is an indication of whether
the candidate is ambiguously or unambiguously identified.

It is advantageous for a user to monitor the
sequence data together with the BLAST results. In view of
these data, it may be apparent that a particular sequence
10 contains excessive sequencing errors. In this case, as
represented by workflow arrow 2609, the user can send a
request (to a laboratory technician) to retrieve the original
stored colony and to perform again amplification and
sequencing steps 2605. The new sequence data is then entered
15 into the system, as indicated by data-flow arrow 2608, and
candidate gene sequences again sought.

During gene identification step 2616, certain
information is returned that it is advantageous to cache.
Look-aside databases 2618 contain this cache. One such look-
20 aside database is a table of accession number synonyms. When
multiple accession numbers are obtained for a candidate
sequence for a fusion insert, they can be stored, along with
the preferred accession number used for gene reference in the
databases of this invention, as synonyms for future look-up.
25 When a further accession number is received, this table can
be searched to determine if it has been encountered
previously, and if so, the corresponding, preferred accession
number used in the databases. Another look-aside database is
a homology database. The results of homology searches can be
30 saved as tables of accession numbers of sequences having
homologies above certain thresholds. For BLAST searches,
such thresholds can be probabilities of e^{-10}, e^{-20}, e^{-30}, e^{-40}, e^{-50}, e^{-60}, e^{-70}, e^{-80}, e^{-90}, or e^{-100}. This table permits doing
simple homology searches efficiently by finding the accession
35 numbers of those sequences having a certain homology with a
query accession number.

5.2.9.2. INTERACTION DATABASE

Based on candidate gene identifications and other information in the identification database, interaction quality control step 2619, on Figure 26, updates interaction
5 database 2620. The interaction database stores two types of information: information on specific protein-protein interactions observed in one or more mating experiments; and information on which colonies containing which interacting fusion inserts were positive in one or more mating
10 experiments (i.e., providing evidence for the protein-protein interaction). In this subsection, the contents and update of this database are described. In the following subsection, the useful ways that this information can be used ("mined") are exemplified.

15   Updates of the interaction database proceed according to at least two general embodiments. Briefly, in a first embodiment, interaction quality control step 2619 formats and presents the data in identification database 2617 relating to each positive colony to a user skilled in biology
20 and, preferably, also skilled in the biology applicable to the type of protein-protein interactions being presented. For a positive colony, the user decides, first, if it is biologically interesting or important, and if so, second, selects from among the candidates those genes, if any, that
25 are actually involved in this interaction. In an alternative embodiment, where the user's decision criteria can be reduced to rules, or to other computer processible representation, the decisions for a colony can be performed automatically by quality control step 2619. See, e.g., Russell et al., 1995,
30 Artificial Intelligence - A Modern Approach, Prentice Hall, chaps 1 and 15, the entirety of this reference is hereby incorporated by reference. Based on these decisions, interaction database 2620 is updated in the following manner. If an interaction between the selected genes is already
35 defined in the database, the new colony information defines in the database a new interacting pair of fusion inserts representing an additional observation of that interaction.

If such an interaction does not yet exist, the new information defines in the database both a new interaction between the selected genes and a new interacting pair representing an observation of that interaction. Also as
5 part of quality control step 2619, the user, or alternatively an automated decision system, can request confirmation tests on the particular positive colony. As represented by workflow arrow 2610, this decision generates requests to perform the tests that displayed, for example, at terminals
10 of the responsible laboratory technicians.

Prior to describing the processing of this step in more detail, the preferable information content of interaction database 2620 is described. The interaction database is conceptually divided into two components. The
15 two components can be represented by physical divisions, by separate groups of tables, by logical views, or by other means known in the art. The first component (the "interaction" component) represents interactions generally, and the second component (the "interacting pair" component)
20 represents interacting pairs evidencing general interactions. An interacting pair from a positive colony evidence an interaction if the fusion inserts observed in that colony are identified as being coded by the genes defining the interaction.
25   In more detail, the interaction component includes information exemplified in Table 1B.

30

35

TABLE 1B: INTERACTION DATABASE - INTERACTIONS

FIELD	DESCRIPTION
gene-1	Identity of gene coding for one interacting protein of pair (database or internal accession number)
description of gene-2	If homology exists name, species origin, tissue origin, and so forth; if no close homology exists, same information for closest homologue
gene-2	Identity of gene coding for other interacting protein of pair (database or internal accession number)
description of gene-1	If homology exists name, species origin, tissue origin, and so forth; if no close homology exists, same information for closest homologue
number of interacting pairs	Total number of positive colonies with interacting fragments from this gene pair
number of independent interacting pairs	Number of colonies with sufficiently different interacting fragments from this gene pair
bi-directional links to interacting pairs	Appropriate links to observed interacting pairs for this interaction (e.g., unique colony identifiers)
interaction type	For example, inhibition/activation of function, direction of interaction, and so forth (determined from biochemical protocols)
interaction source	Interaction observed in this facility, observed in other facility, entered from literature reference, etc.

A protein-protein interaction, according to this embodiment, is considered to occur between two proteins coded for by the two genes described in the fields gene-1 and gene-2. Where interactions can be observed that simultaneously involve three or more proteins, the data structures of the interaction database can be adapted in straightforward ways apparent to those of skill in the art. If already known genes can be identified for the interaction, they are

- 85 -

identified by the preferred sequence database (GenBank) accession number. The description field can contain (as in the identification database) a description of the gene and its function. If coding sequences already known in databases

5 cannot be identified for an observed fusion insert sequence, a distinguishable internal sequence number is generated and associated with the observed insert. Perhaps, for example, the only highly homologous currently known sequence is from the wrong species, e.g., a mouse sequence highly homologous

10 to an insert from a human sample. Alternatively, perhaps no currently known sequence is sufficiently homologous to an insert sequence to be its possible source. Advantageously, when a generated sequence number for a new sequence is used, the description field can point to the most homologous known

15 genes.

Additionally, interaction database fields relate to the observed colonies, or interacting fusion insert pairs, evidencing a general interaction. First, at least, the total number of such interacting pairs is recorded. Second, the

20 total number of "independent" interacting pairs is also recorded. An independent interaction pair is defined as follows: Out of the total number of interacting pairs, it is likely that several will in fact be substantially identical. For example, several observed positive colonies can arise

25 from doublings of a single mated cell, or a single insert from the original DNA colony can be cloned into several different plasmids. Accordingly, two interacting pairs are considered substantially identical if both of their fusion inserts are the same to within expected sequencing errors.

30 Typically, two inserts are identical if they are of approximately the same length (to within less than preferably 5% or 10% of the insert length) and have substantially homologous nucleotide sequences (to within less than preferably 5% or 10% of the number of nucleotides).

35 Otherwise, the two interacting pairs are considered not substantially identical and thus "independent." For example, an insert of a first interacting pair can be different in

that it is: longer and overlaps (as determined by comparing
3' and 5' end information) the gene sequence of an insert of
a second interacting pair, or is displaced with respect to
the second insert, or has substantially different sequence,
5 or so forth.   Using such criteria, a new interacting pair
can be compared to the already recorded interaction pairs for
an interaction in order to determine if it is a new
independent interacting pair.   The greater the number of
independent interacting pairs evidencing an interaction, the
10 more statistically significant the interaction is considered
to be.

         The interaction database also maintains bi-
directional links between general interaction information and
interacting fusion insert pair information.   This can be done
15 explicitly by providing unique identifiers for interactions
and interacting pairs in the database and by storing
identifiers for pairs with the related general interaction
and vice versa.   Alternatively, existing identifiers can be
used to make this bi-directional link.   Unique colony
20 identifiers can point to interacting pair information, and
unique gene accession numbers can point to general
interaction information.

         Finally, interaction information can include
optional associated information, e.g., interaction source and
25 interaction type.   Interaction source indicates where the
observation of this interaction was made.   For example, it is
advantageous for interactions observed in other laboratories
or reported in the literature to be available in the
interaction database.   Such information can be manually
30 entered.   Interaction type provides biochemical information,
if available, on the interaction.

         Next, Table 1C provides more detail on the
interacting pair components.

35

TABLE 1C: INTERACTION DATABASE - INTERACTING PAIRS

FIELD	DESCRIPTION
mating experiment identification	For interaction observed in this facility, appropriate unique identification of the mating experiment with links to description of libraries used, precise protocols, and so forth
colony identification	Appropriate unique identification of the positive colony with this interacting pair (in particular for future retrieval from storage)
gene-1	Identity of gene selected as coding for one interacting protein of pair (database or internal accession number)
3'-5' position-1	Location of fusion fragment on gene (nucleotide positions of fragments ends)
other identifying data	Fragment from activation/binding domain, pointer to measured sequence, and so forth
gene-2	Identity of gene selected as coding for other interacting protein of pair (database or internal accession number)
3'-5' position-2	Location of fusion fragment on gene (nucleotide positions of fragments ends)
other identifying data	Fragment from activation/binding domain, pointer to measured sequence, and so forth
confirmation test data	For example, plasmid drop-out test results, matrix-mating test results, and so forth

Interacting pair information includes the unique experiment and colony identifiers which identify the colony and its generating experiment. It also includes information identifying both fusion inserts of the identified positive colony. The fields gene-1 and gene-2 contain the accession numbers of the gene chosen as the source of the fusion insert. Alternatively, these fields can be labeled as activation domain and binding domain, as is done in the

- 88 -

interaction database (Table 1A). These accession numbers can
be sequence database accession numbers identifying known
genes, or alternatively, internal accession numbers
identifying previously putatively new nucleotide sequences.
5 The location of the insert sequence in the gene selected as
most homologous to the insert is indicated by the 3' and 5'
nucleotide numbers in the gene sequence of each end of the
insert. Further fusion-insert information includes, e.g.,
whether the insert was found in an activation domain or a
10 binding domain library, a pointer to the measured insert
sequence, and so forth. Further information for the colony
includes, e.g., the results of confirmatory tests performed
on the colony, or where further steps 2607 are performed.
The interaction pair data base component can be directly
15 stored or indirectly point to the resulting information.

The components of the interaction database, like
the identification database, are preferably stored in a
relational format using an appropriate normal form.
Appropriate table structures and indices will be apparent to
20 those of skill in the art. In particular, the interactions
and the interacting pairs can be stored in separate tables.
Links between these tables can be implicit, depending on
matching the contents of fields, or explicit, depending on an
additional table storing explicit pointers. Alternatively,
25 this database can be stored according to other database
formats, or less preferably, as user-maintained files.

The remainder of this subsection describes the
processes and methods used by interaction quality control
step 2619 in order to create or to update the interaction
30 database based on information in the identification database.
In view of the information content in these databases, three
decisions for each positive colony are made during this
processing: (1) selection of the genes coding for the
interacting pair, or assignment of an internal accession
35 number or numbers in case one or both genes cannot be
identified; (2) location of the 3' and 5' ends of the fusion
insert on the selected gene; (3) decision as to whether this

interacting pair is independent of already known interacting pairs. In one embodiment of this step, the quality control process retrieves and displays to a user information for positive colonies from the identification database on

5 formatted screens. The user reviews this information, makes the three decisions, and trigger updates to the interaction database accordingly. Advantageously, the user makes these decision largely according to articulated and established rules. Accordingly, in another embodiment, the three

10 decision rules are encoded in a format suitable for a computer-implemented rule based processor (such as one of the expert system packages known in the art or commercially available). The rule processor then makes proposed decisions and proposed updates, which are displayed for user acceptance

15 or revision. In a further embodiment, the three decisions and database update are entirely automated by the rule processor, which the user only later reviews and, perhaps, revises during, e.g., database browsing step 2621.

Accordingly, presented herein is an exemplary set

20 of decision rules for controlled update of the interaction database based on information in the identification database. These rules are capable of being applied by a computer-implemented rule based processor or by an individual user. An exemplary rule for the first decision proceeds, first, by

25 discarding candidate genes that are not from the same species as the source of the library used in making the fusion plasmids. Thus if the positive colony resulted from a mouse derived activation domain library, then only mouse genes are considered further as possibly coding for an activation

30 domain fusion insert. Second, homologies to anti-sense strands, if known, are also discarded. Third, optionally, candidate genes for both fusion inserts are grouped together by the general functions of their encoded proteins. For example, general protein functions can include cell cycle

35 control, intra- or inter-cellular signaling, cell-specific function such as metabolic or synthetic activities, and so forth. Other possible classifications of protein function

- 90 -

will be readily apparent to those of skill in the art.
Proteins of unknown function can be assigned to a default
group that corresponds to every other functional group.   .
Fourth, the homology scores of all pairs of genes with

5 protein of corresponding function is retrieved, and that pair
of corresponding genes with the highest homology scores to
the observed fusion inserts is assigned to the fusion
inserts. The accession numbers of these genes are used to
update the interaction database. Preferably, the selected

10 genes will have homology scores above a certain threshold.
In the case of BLAST search results, such a preferred
homology threshold is between a probability of e^{-5} and e^{-10} for
shorter insert sequences of no more than a few tens (e.g.,
20) of nucleotides, or alternately, in the case of longer

15 insert sequences may be greater than, e.g., e^{-20}, e^{-30}, e^{-40}, e^{-50}, e^{-60}, e^{-70}, e^{-80}, e^{-90}, or e^{-100}. If no such genes exist, the
fusion inserts are considered to be from unknown genes and
internal accession numbers are generated for updating the
interaction database. An alternative rule for this first

20 decision is simply to select, independently for each fusion
insert, the gene with the highest homology score to the sense
strand, if known, that is above a certain threshold and is
from the correct species. If no such gene exists, an
internal accession number is generated.

25   An exemplary rule for the second decision, in the
case of BLAST homology searches, is simply to assign the 3'
and 5' ends of the insert to be the nucleotide numbers of the
subsequence matched from the most homologous gene found by
BLAST. In the case of QEA signals, an alternative rule for

30 finding the 3' and 5' ends has previously been described. An
exemplary rule for the third decision, whether or not the
current insert pair is independent, proceeds by retrieving an
example of each of the independent interacting pairs already
found for this interaction and to compare the degree of

35 homology of both the activation and binding domain fusion
inserts, respectively, with the current pair. A simple test
for insert homology is to check, first, that the 3' and 5'

ends of both inserts differ preferably by less than 1%, or by
less than 5%, or by less than 10%, of the total length of the
insert, and to check, second, whether the number of different
nucleotides is preferably less than 1%, or less than 5%, or
5 less than 10%, of the total number of nucleotides in the
insert. These ranges are to accommodate for expected
sequencing errors. Two inserts falling within these bounds
are taken not to be independent. Alternatively, a homology
search tool (such as BLAST) can be used to estimate the
10 degree of homology between the current inserts and the
retrieved examples. Current inserts without any significant
homology to the retrieved examples are a new independent
pair.

Thereby, interaction quality control step 2619
15 updates interaction database 2620, whose information content
is summarized in Tables 1B and 1C, with new interactions and
new interacting pairs based on information in identification
database 2617 by computer-implemented and rule-based
processing according to, e.g., the previously described rules
20 defining the biological significance of candidate genes. The
interaction database can also be manually updated with
interaction information received from others or reported in
the literature.

25   5.2.9.3. <u>INTERACTION DATABASE FUNCTIONS</u>
The interaction database contains valuable
information that can be usefully accessed and analyzed for
diverse purposes. In this subsection, three particular
analysis functions applicable to this information are
30 presented: database browsing function 2621, protein domain
identification function 2622, and interaction pathway
construction function 2623. However, included within this
invention are the other diverse uses of the valuable protein-
protein interaction information stored in the interaction
35 database that will be apparent to a person skilled in the
art.

Database browsing 2621 is, on the one hand, a general function for selecting (also called "filtering" herein) a subset of information from the database for function processing. On the other hand, database browsing

5   also includes the formatted display of the selected or filtered information on a user's terminal. Especially where the interaction database is stored in relational format, such subset selection can be easily done according to any possible relational query, as is known in the art. The selected

10 subset can be formatted for display by presenting the selected information rows in a labeled, tabular layout with a scrolling capability useful to view further rows and fields not immediately viewable on one screen. Any data in the interaction database, either from the interaction component,

15 the interacting fision insert pair component, or both components combined can be selected and displayed.

In particular embodiments, it is useful to assist the user by providing pre-established queries, or filters, available for easy selection. Figures 28A-B illustrate

20 examples of such useful and hierarchically arranged filters. Figure 28A allows a user to select a particular mating experiment (identified by the column labeled "screen") for further display. The additional columns display further data characterizing each experiment: the "library" columns

25 identify the DNA libraries used to construct the plasmid libraries; the "name" column displays the activation domain and binding domain library identities; the further columns display the laboratory status of the experiment. Once a user selects a particular experiment (or "screen"), Figure 28B

30 illustrates a next filter that allows the user to further select particular information for this experiment from the interaction database. The data can be selected by the status of confirmation tests, by the "screen" type, by the source of the interaction , or by the number of independent interacting

35 pairs (referred to as "isolates" on Figure 28B). The "list" options control the display of selected data. Finally, the "SUBMIT TO PATHMAKER™" permits the selected interactions to

be assembled into pathways by pathway construction function 2623.

In particular, the "screen" filter permits selection of "forward," "reverse," and "bi-directional"
5 screens. For forward screens, data is selected in which a chosen library was used to make activation domain fusions. For reverse screens, the data refers to results in which the library was used to make binding domain fusions. Finally, for "bi-directional" screens, the data refers to results in
10 which the library was used to make both activation domain and binding domain fusions, in the same or in separate mating experiments. For example, an interaction of gene-A and gene-B passing a "bi-directional" filter has both at least one interacting pair in which gene-A is present in an activation
15 domain fusion and gene-B is present in a binding domain fusion and also at least one interacting pair in which gene-A is present in a binding domain fusion and gene-B is present in an activation domain fusion. Finding interactions present in a "bi-directional" screen increases confidence in the
20 biological significance of the interaction, and decreases the possibility that the interaction is merely an artifactual occurrence.

Pathway construction function 2623 automates construction of protein interaction pathways, which represent
25 the links by which proteins can interact with distant proteins through intermediate proteins. Preferably, this function also provides for graphical display of the resulting pathways. Figure 29 illustrates such a graphical display of a simple pathway, in which three proteins, Proteins A, B, and
30 C, have been found to all individually interact with the protein MDM2. These individual, pair-wise interactions create three possible pathways by which Protein A can interact with Protein B or with Protein C and by which Protein B can interact with Protein C, all mediated by
35 protein MDM2.

Determination of such pathways start with selection of a subset of the interactions stored in the interaction

database for pathway analysis. As described with respect to
the database browsing function, this selection can be by a
relational query.   Alternatively, all the interactions in
the database can be analyzed into pathways. Pathway analysis
5 begins with representing the selected protein interactions as
a graph, which is defined by a set of vertices, V, and a set
of edges, E, each edge connecting two vertices. Each vertex
represents one gene, or protein, and the set of vertices V is
assembled by retrieving all the distinct proteins, or genes,
10 present in the selected subset of the interaction database.
Each edge represents one protein-protein interaction since
each such interaction links two genes in the set of vertices,
and the set of edges, E, is assembled by retrieving the set
of selected interactions. For example, a graph for the
15 pathways illustrated in Figure 29 is defined by the set of
vertices (Protein A, Protein B, Protein C, MDM2) and the set
of edges ((Protein A, MDM2, ...), (Protein B, MDM2, ...)
(MDM2, Protein C, ...)) ("..." represents additional
interaction information). Having defined the interaction
20 graph, each separate pathway is represented by a connected
component of this graph. Two vertices are in the same
connected component, if they are connected by a path of
edges. No path of edges connects two vertices in different
connected components. Finding connected components of a
25 graph is well known to those of skill in the art, and can be
done by the basic depth-first search algorithm. See, e.g.,
Sedgewick, 1990, <u>Algorithms In C</u>, Addison-Wesley Publishing
Co., chap. 29, the entirety of this reference is incorporated
herein by reference.
30   Finally, each connected component is then
separately formatted and displayed on a user's computer
screen. For ease of viewing, the graph is preferably
displayed with the protein, or gene, vertices well separated
on the screen, and also preferably, if possible, with the
35 edges, representing interactions not crossing (that is as a
planar graph). Since such a display can be difficult to
create in general, an exemplary approximation is to place

graph vertices on the screen according to a simulated
annealing algorithm, which approximately minimizes an
"energy" function using statistical techniques. See, e.g.,
Press, et al., 1988, Numerical Recipes in C, Cambridge
5 University Press, Cambridge, U.K., which is herein
incorporated by reference in its entirely. The preferred
display goals are approximately achieved by minimization of
an "energy" function, which grows large both when two
vertices are close and also when edges cross. An exemplary
10 such function includes a term for each vertex that depends on
the inverse of the distance to the nearest neighbor of that
vertex as well as a large positive factor for each edge
crossing. Simulated annealing then successively perturbs
vertex screen placement in order to search for a placement
15 approximately minimizing the energy function.

        Further preferable display features include coding
gene information in the appearance of its vertex by, e.g.,
the vertex color, or coding interaction information by, e.g.,
the edge color or graphic, and so forth. Additional
20 information, beyond that so coded, on a gene or an
interaction can be obtained by "clicking" on their screen
representations. For example, clicking on a edge
representing an interaction can call up a window in which
summary or graphical information on the interacting pairs
25 evidencing that interaction is presented. Such information
can include a graphical representation of where the fusion
inserts are located on the coding sequence of the gene.

        Finally, domain identification function 2622
automates locating the actual protein domains responsible for
30 an interaction. In a first simple embodiment, applicable to
a single pair-wise interaction, for example that of Protein A
with protein MDM2, the locations of all the fusion inserts on
the gene sequence are simply intersected in order to obtain a
location common to and included in all the fusion inserts.
35 The protein domain responsible for the interaction evidenced
by these fusion inserts lies within the amino acid sequence
coded by this common region. Figure 30 illustrates this

processing. Sequence 3001 represent the entire gene coding
sequence for the one interacting protein participating in an
interaction. Sequences 3002, 3003, and 3004 represent three
fusion inserts fragments from this gene that were found in
5 three independent interacting pairs evidencing this
interaction. They are illustrated aligned between their 3'
and 5' ends as determined in previous processing steps.
Subsequence 3005 of sequence 3001 is the intersection of the
three inserts. Clearly, the protein domain responsible for
10 the interaction must be encoded by (all or perhaps a portion
of) the subsequence 3005, as this is the only common amino
acid sequence to all the interacting protein fragments.
Subsequence 3005, as illustrated, can be computed as the
sequence lying between a 3' boundary, which is the minimum of
15 the 3' ends of all the fusion inserts, and a 5' boundary,
which is the maximum of all the 5' ends of the fusion
inserts. Only inserts from independent interacting pairs
need be retrieved for this determination.

        Domain identification is more certain if the same
20 domain is found in a bi-directional screen, when the inserts
from the protein are fused with both activation domains and
binding domains. Domain identification is also more certain
if known motifs can be identified in the domain. After
domain location is determined, the amino acid sequences
25 encoded can be searched for known motifs.

        In a further embodiment, additional domain
information can be obtained in certain cases. By way of
example, referring to Figure 29, the ternary interaction of
Protein A and Protein B intermediated by MDM2 can provide
30 additional domain information according to the following
procedure. First, intersection domains are determined as
previously described for Protein A and MDM2 and for Protein B
and MDM2. If both Proteins A and B interact with the same or
overlapping MDM2 domain, then more information may be
35 obtained by comparing the domains found in Proteins A and B
as follows. A BLAST comparison of these two domains may
reveal homologous structures of a probability which might be

ignored if the functional relationship revealed by the
interaction were not a priori known. The domains may be
compared by protein search tools, especially search tools
capable of evaluating multiple alignment between the two
5 domains, in order to reveal structural relationships, the
presence of motifs, and so forth, at the amino acid sequence
level. Further, other techniques for extracting domain
information from binary, ternary, and higher-order
interactions will be apparent to those of skill in the art.
10 Such techniques are within the applications of the
interaction database of this invention.

        The information-processing aspects of this
invention also include those variations and elaborations that
are apparent to those of skill in the art in view of the
15 disclosure herein. In particular, the experimental data and
workflow controls can be extended to manage the additional
steps of mating experiments prior to fusion protein
characterization or after confirmation tests. Automation of
screening interaction agonists and antagonists is an
20 especially advantageous extension.

            5.3.   INTEGRATED ISOLATION OF INHIBITORS
                   OF AN INTERACTIVE POPULATION

25   The present invention also provides methods for
identifying inhibitors or enhancers of protein-protein
interactions. The method of identifying inhibitors provided
by the invention provides for greater ease and higher
throughput than prior art methods, inter alia, through the
30 ability to select for inhibitors based on cell survival. The
present invention is particularly valuable in that it enables
one to identify not only the interacting proteins that are
unique to or characteristic of a particular situation, but
also enables the identification of inhibitors of such
35 interactions. The invention provides a method of detecting
an inhibitor of a protein-protein interaction comprising (a)
incubating a population of cells, said population comprising

cells recombinantly expressing a pair of interacting
proteins, said pair consisting of a first protein and a
second protein, in the presence of one or more candidate
molecules among which it is desired to identify an inhibitor

5   of the interaction between said first protein and said second
protein, in an environment in which substantial death of said
cells occurs (i) when said first protein and second protein
interact, or (ii) if said cells lack a recombinant nucleic
acid encoding said first protein or a recombinant nucleic

10 acid encoding said second protein; and (b) detecting those
cells that survive said incubating step, thereby detecting
the presence of an inhibitor of said interaction in said
cells. In a preferred aspect, the population of cells
comprises a plurality of cells, each cell within said

15 plurality recombinantly expressing a different said pair of
interacting proteins. In various embodiments, the plurality
of cells consists of at least 10, at least 100, or at least
1000 cells (corresponding to different pairs of interacting
proteins being assayed in a single assay). In a preferred

20 embodiment, the pair(s) of interacting proteins in the cells
being assayed consist of a first fusion protein and a second
fusion protein, each said first fusion protein comprising a
first protein sequence and a DNA binding domain; each said
second fusion protein comprising a second protein sequence

25 and a transcriptional activation domain of a transcriptional
activator; and in which the cells contain a first nucleotide
sequence operably linked to a promoter driven by one or more
DNA binding sites recognized by said DNA binding domain such
that an interaction of said first fusion protein with said

30 second fusion protein results in increased transcription of
said first nucleotide sequence, and in which the cells are
incubated in an environment in which substantial death of the
cells occurs (i) when increased transcription occurs of the
first nucleotide sequence or (ii) if the cells lack a

35 recombinant nucleic acid encoding the first fusion protein or
a recombinant nucleic acid encoding the second fusion
protein. The cells in which the assay is carried out are

preferably (but need not be) yeast cells, which can be
haploid or diploid.

In a specific embodiment, an assay for the presence
of an interacting protein pair is carried out as described in
5 the sections *supra*, except that it is done in the presence of
one or more candidate molecules which it is desired to screen
for the ability to affect an interaction between a protein-
protein pair that results in transcription from the Reporter
Gene. An increase or decrease in Reporter Gene activity
10 relative to that present when the one or more candidate
molecules are absent indicates that the candidate molecule
has an effect on the interacting pair. For example, a
decrease in (e.g., absence of) Reporter Gene activity that
would otherwise occur in the absence of a candidate molecule,
15 due to the presence of an interacting pair, indicates that
the candidate molecule is an inhibitor of the interaction
exhibited by the protein pair. In a preferred embodiment,
selection of positive interactants (colonies) is carried out;
these colonies are exposed to candidate inhibitor molecule(s)
20 and are selected again, this time for lack of interaction
(e.g., by selection for survival in medium containing 5-FOA
wherein *URA3* is a Reporter Gene, or by selection for survival
in medium containing α-amino-adipate wherein *LYS2* is a
Reporter Gene, or the other methods of negative selection
25 described in Section 5.1 above; selection of cells that do
not display a signal generated by a Reporter Gene (e.g., in
the case of *lacZ*, by activity on the β-gal substrate X-gal
(5-bromo-4-chloro-3-indolyl-β-D-galactoside)). The
environment in which selection is carried out preferably also
30 selects for the presence of the recombinant nucleic acids
encoding the interacting pair of proteins. Thus, for
example, the proteins are expressed from plasmids also
expressing a selectable marker, thus facilitating this
selection.

35   For detecting an inhibitor, candidate inhibitor
molecules can be directly provided to a cell containing an
interacting pair, or, in the case of candidate protein

- 100 -

inhibitors, can be provided by providing their encoding nucleic acids under conditions in which the nucleic acids are recombinantly expressed to produce the candidate proteins within the cell. The recombinantly expressed candidate

5 inhibitors preferably comprise a nuclear localization signal to facilitate their import into the nucleus and exposure to the interacting protein pair.

A preferred exemplary method for detecting the presence of inhibitors of protein-protein interactions is

10 shown in Figure 6. The interactive population is grown in a 96-well format with each well containing 200 μl of media. If the two interacting proteins are plasmid-borne then the media preferably selects for maintenance of the plasmids, e.g., the media lacks those markers, like tryptophan or leucine that

15 allow selection for the plasmids bearing TRP1 or LEU2, respectively. (This maintenance of selective pressure is obviated if the genes encoding the two proteins are not plasmid-borne but have been integrated into the chromosome instead). Each well contains all the colonies that were

20 identified as containing protein interactants from an N x M assay of protein interactions according to the invention. Thus, each well is representative of all the interactive proteins present in a particular population. In the preferred embodiment of the invention, the Reporter Gene used

25 for selection of interaction and selection of inhibition of interaction is the URA3 gene. Interaction between the two fusion proteins causes the yeast to grow in the absence of uracil, allowing selection of the interacting colonies. However, activation of the URA3 gene causes the yeast to die

30 in medium containing the chemical 5-fluoroorotic acid (5-FOA; (Rothstein, 1983, Meth. Enzymol. 101:167-180)). After a growth period that is sufficient for early log-phase growth (a cell density of about 1 x 10^7 cells/ml), the cells are exposed to inhibitor(s) for 1-2 hours. Then an appropriate

35 dilution of the cells is transferred to a 96-well plate containing 200 μl media lacking uracil to activate the transcription of the URA3 gene as a result of interaction

between the two hybrid proteins in the presence of inhibitor(s). After this, an appropriate dilution of the cells is transferred to a 96-well plate containing 200 ml media made up of 5-FOA and the inhibitor(s). At this step

5 an alternative is to transfer 1 μl onto a 96-slot grid on solid media containing 5-FOA and the inhibitor(s) at the desired concentration.

Growth will be evident only in those instances where inhibition of the protein-protein interaction occurs.

10 As a preferred control, all the cells should be able to grow in the absence of 5-FOA but in the presence of the inhibitor. Thus, in a single screen, the inhibitor and the pair of interacting proteins it inhibits are identified. The identities of the interacting proteins that are inhibited are

15 revealed by characterizing the genes that encode these interacting proteins.

The presence of more than one inhibited pair in a well would be indicated, e.g., by sequence analysis. In such an instance, the cells surviving in the presence of 5-FOA can

20 be diluted, and the inhibition assay repeated. Ultimately, the cells are diluted and streak-purified so as to isolate single colonies representing a single pair of interacting proteins. Then the inhibition assay is repeated on these streak-purified isolates.

25 In the 96-well format of this assay, the activity of a *lacZ* Reporter Gene can also be assayed enzymatically. The activity of the *lacZ* gene can be determined by assaying the β-galactosidase levels. This can be done in a high throughput fashion as chemiluminescent assays or fluorescent

30 assays using substrates that are chemiluminescent (Jain and Magrath, 1991, Anal. Biochem. 199:119-124) or fluorescent (Fluoreporter lacZ/β-galactosidase quantitation kit from Molecular Probes Inc.).

Use of a Reporter Gene that encodes a selectable

35 marker (e.g., *URA3* or *LYS2*) that can be negatively selected against is preferred over the sole use of a Reporter Gene that encodes a detectable marker (e.g., *lacZ*), since negative

selection for a selectable marker can be carried out on each
of multiple interacting pairs within a single well, thus
allowing "multiplex" analysis (analysis of pools of cells·
containing interacting pairs in one well), thus increasing
5 throughput. This is because in the use of negative
selection, survival of any cells indicates that at least one
inhibited pair is present; in contrast, lack of detection of
a detectable marker occurs only if all interacting pairs in
the well are inhibited, while detection of a detectable
10 marker indicates that at least one interacting pair in the
well is not inhibited but does not indicate whether or not
any of the other potential pairs present are inhibited.

This embodiment of the invention is well suited to
screen chemical libraries for inhibitors of protein-protein
15 interactions.

Exemplary libraries are commercially available from
several sources (ArQule, Tripos/PanLabs, ChemDesign,
Pharmacopoeia). In some cases, these chemical libraries are
generated using combinatorial strategies that encode the
20 identity of each member of the library on a substrate to
which the member compound is attached, thus allowing direct
and immediate identification of a molecule that is an
effective inhibitor. Thus, in many combinatorial approaches,
the position on a plate of a compound specifies that
25 compound's composition. Also, in one example, a single plate
position may have from 1-20 chemicals that can be screened by
administration to a well containing the interactions of
interest. Thus, if positive inhibition is detected, smaller
and smaller pools of interacting pairs can be assayed for
30 inhibition. By such methods, many inhibitors can be screened
against many interactors (see, e.g., Figure 6).

Many diversity libraries suitable for use are known
in the art and can be used to provide compounds to be tested
as inhibitors according to the present invention.
35 Alternatively, libraries can be constructed using standard
methods. Chemical (synthetic) libraries, recombinant

expression libraries, or polysome-based libraries are
exemplary types of libraries that can be used.

      The libraries can be constrained or semirigid
(having some degree of structural rigidity), or linear or
5 nonconstrained. The library can be a cDNA or genomic
expression library, random peptide expression library or a
chemically synthesized random peptide library. Expression
libraries are introduced into the cells in which the
inhibition assay occurs, where the nucleic acids of the
10 library are expressed to produce their encoded proteins.

      In one embodiment, the peptide libraries used in
the present invention may be libraries that are chemically
synthesized *in vitro*. Examples of such libraries are given
in Houghten et al., 1991, Nature 354:84-86, which describes
15 mixtures of free hexapeptides in which the first and second
residues in each peptide were individually and specifically
defined; Lam et al., 1991, Nature 354:82-84, which describes
a "one bead, one peptide" approach in which a solid phase
split synthesis scheme produced a library of peptides in
20 which each bead in the collection had immobilized thereon a
single, random sequence of amino acid residues; Medynski,
1994, Bio/Technology 12:709-710, which describes split
synthesis and T-bag synthesis methods; and Gallop et al.,
1994, J. Medicinal Chemistry 37(9):1233-1251. Simply by way
25 of other examples, a combinatorial library may be prepared
for use, according to the methods of Ohlmeyer et al., 1993,
Proc. Natl. Acad. Sci. USA 90:10922-10926; Erb et al., 1994,
Proc. Natl. Acad. Sci. USA 91:11422-11426; Houghten et al.,
1992, Biotechniques 13:412; Jayawickreme et al., 1994, Proc.
30 Natl. Acad. Sci. USA 91:1614-1618; or Salmon et al., 1993,
Proc. Natl. Acad. Sci. USA 90:11708-11712. PCT Publication
No. WO 93/20242 and Brenner and Lerner, 1992, Proc. Natl.
Acad. Sci. USA 89:5381-5383 describe "encoded combinatorial
chemical libraries," that contain oligonucleotide identifiers
35 for each chemical polymer library member. Compounds
synthesized so as to be immobilized on a substrate are

- 104 -

released from the substrate prior to use in the inhibition assay.

Further, more general, structurally constrained, organic diversity (e.g., nonpeptide) libraries, can also be
5 used. By way of example, a benzodiazepine library (see e.g., Bunin et al., 1994, Proc. Natl. Acad. Sci. USA 91:4708-4712) may be used.

Conformationally constrained libraries that can be used include but are not limited to those containing
10 invariant cysteine residues which, in an oxidizing environment, cross-link by disulfide bonds to form cystines, modified peptides (e.g., incorporating fluorine, metals, isotopic labels, are phosphorylated, etc.), peptides containing one or more non-naturally occurring amino acids,
15 non-peptide structures, and peptides containing a significant fraction of γ-carboxyglutamic acid.

Libraries of non-peptides, e.g., peptide derivatives (for example, that contain one or more non-naturally occurring amino acids) can also be used. One
20 example of these are peptoid libraries (Simon et al., 1992, Proc. Natl. Acad. Sci. USA 89:9367-9371). Peptoids are polymers of non-natural amino acids that have naturally occurring side chains attached not to the alpha carbon but to the backbone amino nitrogen. Since peptoids are not easily
25 degraded by human digestive enzymes, they are advantageously more easily adaptable to drug use. Another example of a library that can be used, in which the amide functionalities in peptides have been permethylated to generate a chemically transformed combinatorial library, is described by Ostresh et
30 al., 1994, Proc. Natl. Acad. Sci. USA 91:11138-11142).

The members of the libraries that can be screened according to the invention are not limited to containing the 20 naturally occurring amino acids. In particular, chemically synthesized libraries and polysome based libraries
35 allow the use of amino acids in addition to the 20 naturally occurring amino acids (by their inclusion in the precursor pool of amino acids used in library production). In specific

embodiments, the library members contain one or more non-natural or non-classical amino acids or cyclic peptides. Non-classical amino acids include but are not limited to the D-isomers of the common amino acids, α-amino isobutyric acid,

5   4-aminobutyric acid, Abu, 2-amino butyric acid; γ-Abu, ε-Ahx, 6-amino hexanoic acid; Aib, 2-amino isobutyric acid; 3-amino propionic acid; ornithine; norleucine; norvaline, hydroxyproline, sarcosine, citrulline, cysteic acid, t-butylglycine, t-butylalanine, phenylglycine,

10 cyclohexylalanine, ß-alanine, designer amino acids such as ß-methyl amino acids, Cα-methyl amino acids, Nα-methyl amino acids, fluoro-amino acids and amino acid analogs in general. Furthermore, the amino acid can be D (dextrorotary) or L (levorotary).

15   A specific embodiment of this invention uses mutant strains of yeast that have a mutation in at least one gene coding for a cell wall component, thereby having modified cell walls that are more permeable to exogenous molecules than are wild-type cell walls, thus facilitating the entry of

20 chemicals into the cell, and rendering such yeast cells preferred for an inhibition assay in which exogenous candidate inhibitor compounds are provided directly to the cell. In one embodiment, mutations in the gene *KNR4* in *Saccharomyces cerevisiae* cause the cell wall to be more

25 permeable to chemicals like X-gal, while not affecting general growth (Hong et al., 1994, Yeast 10:1083-1092). The reporter strains are made mutant with respect to gene *KNR4* to facilitate entry of inhibitor compounds. Similarly, in other embodiments, mutations in genes that influence the cell wall

30 integrity (reviewed in Stratford, 1994, Yeast 10:1741-1752) are incorporated into the reporter strain so as to make the cell wall more permeable.

          In a specific embodiment of the invention, the prospective inhibitors are peptides that are genetically

35 encoded and either plasmid-borne or are introduced into the chromosome through homologous recombination. The peptides to be screened are thus provided by recombinant expression

within the cell in which the inhibition assay occurs. The peptides are preferably expressed as a fusion to a nuclear localization sequence. The interactive population (preferably the entire population) from an M x N screen is
5 pooled together and then transformed with a library of plasmids encoding peptides to be tested as potential inhibitors. Alternatively, genes encoding the peptides are introduced directly into the chromosome by first cloning the genes into an integration plasmid containing the yeast
10 sequences that donate the site necessary for homologous recombination. The transformed yeast cells are then plated on media that selects for inhibition events. In the preferred embodiment of the invention, the reporter gene for interaction and inhibition of the interaction will be the
15 URA3 gene. Thus, transformants that emerge in media containing 5-FOA represent peptide inhibitors that inhibit specific protein-protein interactions.

In another embodiment, DNA from a microorganism that reconstitutes synthetic pathways for a compound (see
20 Hutchinson, 1994, Bio/Technology 12:375-380; Alvarez et al., 1996, Nature Biotechnology 14:335-338) can be introduced into the cell in which the inhibition assay takes place, so as to be recombinantly expressed by the cell such that the compound is synthesized within the cell. If the synthesized compound
25 blocks the protein interactants, such cells containing an inhibitor of the interacting pair can be detected by methods as described above. By sequencing the DNA in the cells in which inhibition of the interactants has thus occurred, a novel inhibitory compound can be identified.
30 The identities of the peptide inhibitors are deciphered by the isolation and sequencing of the plasmids that encode these peptides. The identities of the pair of interacting proteins, whose interaction has been inhibited by the peptide, are identified by isolation and sequencing the
35 plasmids that encode these two proteins. The sequences of the inhibitor peptide and those of the interacting proteins can also be obtained by amplifying the protein and peptide

encoding region by PCR or other methods and sequencing of the same. Specific primers can be used to amplify the peptide or the DNA-binding fusion protein or the activation domain fusion protein.

5   In a particular embodiment of the invention, cells are incubating in the presence of candidate inhibitor molecules by expressing such molecules within the cell from recombinant nucleic acids comprising the following operably linked components (a) an ADC1 promoter; (b) a nucleotide
10 sequence encoding a candidate molecule fused to a nuclear localization signal; and (c) an ADC1 transcription termination signal (see e.g., Section 6.8). In a particular embodiment, the candidate molecules are expressed from purified expression vectors comprising the following
15 components: (a) a promoter active in yeast; (b) a first nucleotide sequence encoding a peptide of 20 or fewer amino acids fused to a nuclear localization signal, said first nucleotide sequence being operably linked to the promoter; (c) a transcription termination signal active in yeast,
20 operably linked to said first nucleotide sequence; (d) means for replicating in a yeast cell; (e) means for replicating in E. coli; (f) a second nucleotide sequence encoding a selectable marker for selection in a yeast cell, operably linked to a transcriptional promoter and transcription
25 termination signal active in yeast; and (g) a third nucleotide sequence encoding a selectable marker for selection in E. coli, operably linked to a transcriptional promoter and transcription termination signal active in E. coli. The means for inserting is preferably one or more
30 suitably located restriction endonuclease recognition sites; the means for replicating in a yeast cell can be any suitable origin of replication; the means for replicating in E. coli can be any suitable origin of replication. The invention provides expression vectors which can be used for expression
35 of candidate inhibitor molecules, such as a purified expression vector comprising the following components: (a) an ADC1 promoter; (b) a first nucleotide sequence encoding a

nuclear localization signal, operably linked to the promoter;
(c) means for inserting a DNA sequence into the vector in
such a manner that a protein encoded by the DNA sequence is
capable of being expressed as part of a fusion protein
5 containing the nuclear localization signal; (d) an ADC1
transcription termination signal, operably linked to the
first nucleotide sequence; (e) means for replicating in a
yeast cell; (f) means for replicating in *E. coli*; (g) a
second nucleotide sequence encoding a selectable marker for
10 selection in a yeast cell, operably linked to a
transcriptional promoter and transcription termination signal
active in yeast; and (h) a third nucleotide sequence encoding
a selectable marker for selection in *E. coli*, operably linked
to a transcriptional promoter and transcription termination
15 signal active in *E. coli*.

5.4. THE QEA™ METHOD
5.4.1. QUANTITATIVE EXPRESSION ANALYSIS METHOD, GENERALLY

According to a Quantitative Expression Analysis
20 ("QEA™") method, to uniquely identify an expressed gene
sequence, full or partial, and many components of genomic
DNA, it is not necessary to determine actual, complete
nucleotide sequences of samples. Full sequences provide far
more information than is needed to merely classify or
25 determine a gene according to the QEA™ method.

In a QEA™ method, expressed sequences are
recognized by codes which are constructed from signals which
represent the presence of short nucleic acid (preferably DNA)
subsequences (hereinafter called "target subsequences") in
30 the sample sequence and include a representation of the
length along the sample sequence between adjacent target
subsequences. The presence of these subsequences is
recognized by subsequence recognition means, including, but
not limited to, restriction endonucleases (hereinafter called
35 "REs"), DNA binding proteins, and oligomers ("probes")
hybridizable to DNA made of, for example, peptide nucleic
acids (hereinafter called "PNAs") (See, *e.g.*, Egholm et al.,

1993, *Nature* <u>365</u>:566-67) or DNAs. The subsequence
recognition means allow recognition of specific DNA
subsequences by the ability to specifically bind to or react
with such subsequences. A QEA™ method, and particularly its
5 computer methods, are adaptable to any subsequence
recognition means available in the art. Acceptable
subsequence recognition means preferably precisely and
reproducibly recognize target subsequences and generate a
recognition signal of adequate signal to noise ratio for all
10 genes, however rare, in a sample, and can also provide
information on the length between target subsequences.

In some QEA™ embodiments, the presence of target
subsequences is directly recognized by direct subsequence
recognition means, including, but not limited to, REs and
15 other DNA binding proteins, which bind and/or react with
target subsequences, and oligomers of, for example, PNAs or
DNAs, which hybridize to target subsequences. In other
embodiments, the presence of effective target subsequences is
recognized indirectly as a result of applying protocols, such
20 as a SEQ-QEA™ method, or *e.g.*, involving multiple DNA binding
proteins together with hybridizing oligomers. In this latter
case, each of the multiple proteins or oligomers recognizes a
separate subsequence and an effective target subsequence is
the combination of the separate subsequences. A preferable
25 combination is subsequence concatenation in the situation
where all the separately recognized subsequences are
adjacent. Such effective target subsequences can have
advantageous properties not achievable by, for example, REs
or PNA oligomers alone. However, the QEA™ method, and
30 particularly its computer methods, are adaptable to any
acceptable subsequence recognition means available in the
art. The computer implemented analysis and design methods
treat target subsequences and effective target subsequences
in the same manner.
35   The signals contain representations of target
subsequence occurrences and a representation of the length
between target subsequence occurrences. In various

embodiments of the QEA™ method these representations may differ. In embodiments where the target subsequences are exactly recognized, as where REs are used, subsequence representation may simply be the actual identity of the

5 subsequences. In other embodiments where subsequence recognition is less exact, as where short oligomers are used, this representation may be "fuzzy". It may, for example, consist of all subsequences which differ by one nucleotide from the target, or some other set of possible subsequences,

10 perhaps weighted by the probability that each member of the set is the actual subsequence in the sample sequence. Further, the length representation may depend on the separation and detection means used to generate the signals. In the case of electrophoretic separation, the length

15 observed electrophoretically may need to be corrected, perhaps up to 5 to 10%, for mobility differences due to average base composition differences or due to effects of any labeling moiety used for detection. As these corrections may not be known until target sequence recognition, the signal

20 may contain the electrophoretic length in base pairs (hereinafter called "bp") and not the true physical length in bp. For simplicity and without limitation, in most of the following description unless otherwise noted the signals are presumed to represent the information conveyed exactly, as if

25 generated by exact recognition means and error or bias free separation and detection means. However, in particular embodiments, target subsequences may be represented in a fuzzy fashion and length, if present, with separation and detection bias present.

30   Target subsequences recognized are typically contiguous. This is required for all known REs. However, oligomers recognizing discontinuous subsequences can be used and can be constructed by inserting degenerate nucleotides in any discontinuous region.

35   A QEA™ method is adaptable to analyzing any DNA sample for which exists an accompanying database listing possible sequences in the sample. More generally, a QEA™

method is adaptable to analyzing the sequences of any biopolymer, built of a small number of repeating units, whose naturally occurring representatives are far fewer that the number of possible, physical polymers and in which small

5 subsequences can be recognized. Thus, it is applicable to not only naturally occurring DNA polymers but also to naturally occurring RNA polymers, proteins, glycans, etc. Typically and without limitation, however, a QEA™ method is applied to the analysis of cDNA samples from any *in vivo* or

10 *in vitro* sources.

A QEA™ method probes a sample with recognition means, the recognition means generating signals, a preferred signal being a triple comprising an indication of the presence of a first target subsequence, an indication of the

15 presence of a second target subsequence, and a representation of the length between the target subsequences in the sample nucleic acids sequence. Each pair of target subsequences may occur more than once in a sample nucleic acid, in which case the associated lengths are between adjacent target

20 subsequence occurrences.

The QEA™ method is preferred for classifying and determining sequences in cDNA mixtures, but is also adaptable to samples with only one sequence. It is preferred for mixtures because it affords the relative advantage over prior

25 art methods that cloning of sample nucleic acids is not required. Typically, enough distinguishable signals are generated from pairs of target subsequences to recognize a desired sequence in a sample mixture. For example, first, any pair of target subsequences may hit more than once in a

30 single DNA molecule to be analyzed, thereby generating several signals with differing lengths from one DNA molecule. Second, even if the pair of target subsequences hits only once in two different DNA molecules to be analyzed, the lengths between the hits may differ and thus distinguishable

35 signals may be generated.

The target subsequences used in the QEA™ method are preferably optimally chosen by computer methods from DNA

sequence databases containing sequences likely to occur in
the sample to be analyzed. Efforts of the Human Genome
Project in the United States, efforts abroad, and efforts of
private companies in the sequencing of the human genome
5 sequences, both expressed and genetic, are being collected in
several available databases (listed in Section 5.4.2).

In a specific embodiment, a QEA™ method comprises
(a) probing a sample comprising a plurality of nucleic acids
having different nucleotide sequences with one or more
10 recognition means, each recognition means causing recognition
of a target nucleotide subsequence or a set of target
nucleotide subsequences; (b) generating one or more signals
from said sample probed by said recognition means, each
generated signal arising from a nucleic acid in said sample
15 and comprising a representation of (i) the identities of
effective subsequences, each said effective subsequence
comprising a said target nucleotide subsequence, or the
identities of sets of effective subsequences, each said set
having member effective subsequences each of which comprises
20 a different one or more member target nucleotide subsequences
from one of said sets of target nucleotide sequences, and
(ii) the length between occurrences of effective subsequences
in said nucleic acid or between one occurrence of one
effective subsequence and the end of said nucleic acid; and
25 (c) searching a nucleotide sequence database to determine
sequences that match or the absence of any sequences that
match said one or more generated signals, said database
comprising a plurality of known nucleotide sequences of
nucleic acids that may be present in the sample, a sequence
30 from said database matching a generated signal when the
sequence from said database has both (i) the same length
between occurrences of effective subsequences or the same
length between one occurrence of one effective subsequence
and the end of the sequence as is represented by the
35 generated signal, and (ii) the same effective subsequences as
are represented by the generated signal, or effective
subsequences that are members of the same sets of effective

- 113 -

subsequences as are represented by the generated signal,
whereby said one or more nucleic acids in said sample are
identified, classified, or quantified. In a preferred
embodiment, the method comprises (a) digesting said sample
5 with one or more restriction endonucleases, each said
restriction endonuclease recognizing a subsequence
recognition site and digesting DNA at said recognition site
to produce fragments with 5' overhangs; (b) contacting said
produced fragments with shorter and longer
10 oligodeoxynucleotides, each said shorter oligodeoxynucleotide
hybridizable with a said 5' overhang and having no terminal
phosphates, each said longer oligodeoxynucleotide
hybridizable with a said shorter oligodeoxynucleotide; (c)
ligating said longer oligodeoxynucleotides to said 5'
15 overhangs on said fragments to produce ligated DNA fragments;
(d) extending said ligated DNA fragments by synthesis with a
DNA polymerase to produce blunt-ended double stranded DNA
fragments; (e) amplifying said blunt-ended double stranded
DNA fragments by a method comprising contacting said blunt-
20 ended double stranded DNA fragments with a DNA polymerase and
primer oligodeoxynucleotides, each said primer
oligodeoxynucleotide having a sequence comprising that of one
of the longer oligodeoxynucleotides; (f) determining the
length of the amplified DNA fragments produced in step (e);
25 and (g) searching a DNA sequence database, said database
comprising a plurality of known DNA sequences that may be
present in the sample, for sequences matching one or more of
said fragments of determined length, a sequence from said
database matching a fragment of determined length when the
30 sequence from said database comprises recognition sites of
said one or more restriction endonucleases spaced apart by
the determined length, whereby DNA molecules in said sample
are identified, classified, or quantified.

       A QEA™ method can be conducted in a "query" mode or
35 a "tissue" mode. "In a QEA™ method "query mode" experiment,
the focus is on determining the expression of several genes,
perhaps 1-100, of interest and of known sequence. A minimal

number of target subsequences is chosen to generate signals,
with the goal that each of the several genes is discriminated
by at least one unique signal, which also discriminates it
from all the other genes likely to occur in the sample. In
5 other words, the experiment is designed so that each gene
generates at least one signal unique to it (a "good" gene,
see *infra*). In a QEA™ method "tissue mode" experiment, the
focus is on determining the expression of as many as
possible, preferably a majority, of the genes in a sample,
10 without the need for any prior knowledge or interest in their
expression. Target subsequences are optimally chosen to
discriminate the maximum number of sample DNA sequences into
classes comprising one or preferably at most a few sequences.
Signals are generated and detected as determined by the
15 threshold and sensitivity of a particular experiment. Some
important determinants of threshold and sensitivity are the
initial amount of mRNA and thus of cDNA, the amount of
molecular amplification performed during the experiment, and
the sensitivity of the detection means. Preferably, enough
20 signals are produced and detected so that the QEA™ computer
methods can uniquely determine the expression of a majority,
or more preferably most, of the genes expressed in a tissue.

QEA™ method signals are generated by methods
utilizing recognition means that include, but are not limited
25 to, REs in a preferred RE/ligase method or in a method
utilizing a removal means, preferably contacting streptavidin
linked to a solid phase with biotin-labeled DNA, for removal
of unwanted DNA fragments.

A preferred embodiment of an RE/ligase QEA™ method
30 is as follows. The method employs recognition reactions with
a pair (or more) of REs which recognize target subsequences
with high specificity and cut the sequence at the recognition
sites leaving fragments with sticky ends characteristic of
the particular RE. To each sticky end, special primers are
35 ligated which are distinctively labeled with fluorochromes
identifying the particular RE making the cut, and thus the
particular target subsequence. A DNA polymerase is used to

- 115 -

form blunt-ended DNA fragments. The labeled fragments are then PCR amplified using the same special primers a number of times, preferably just sufficient to detect signals from all sequences of interest while making relatively small signals
5 from the linearly amplifying singly cut fragments. The amplified fragments are then separated by length using electrophoresis, and the length and labeling of the fragments is optically detected. In order to improve the quality of the QEA™ method signals, it is preferable to conjugate a
10 capture moiety with one or more of the primers and then to separate unwanted reaction products by a method comprising contacting the reaction products with a binding partner of the capture moiety, washing away unbound products, and then separating by length those single strands which are denatured
15 from the bound products. See Sec. 6.1.12.2.1 ("QEA™ Method Preferred For Use In A SEQ-QEA™ Method"). Optionally, single stranded fragments can be removed by a binding hydroxyapatite, or other single strand specific, column or by digestion by a single strand specific nuclease. Also, the
20 QEA™ method is adaptable to other functionally equivalent amplification and length separation means. In this manner, the identity of the REs cutting a fragment, and thereby the subsequences present, as well as the length between the cuts is determined.

25   In an exemplary QEA™ method utilizing a removal means, which has improved quantitative characteristics and is also adapted to highly sensitive detection systems, cDNA is amplified using at least one internally biotinylated primer. The cDNA is then cyclized, cut with a pair of REs, and
30 specifically labeled primers are ligated to the cut ends, as discussed in § 5.4.3.2 (entitled "Second Alternative RE Embodiment"). The singly cut ends attached to the biotinylated synthesis primers are removed with streptavidin or avidin beads leaving highly pure labeled double cut cDNA
35 fragments without any singly cut and labeled background fragments. With a sufficiently sensitive optical detection system, these pure doubly cut and labeled fragments can be

separated by length (e.g., by electrophoresis or column
chromatography) and directly detected without amplification.
If amplification is needed, absence of the DNA singly cut
fragment background improves signal to noise ratio permitting
5 fewer amplification steps and, thereby, decreased PCR
amplification bias.

Optional alternatives can provide increased
discrimination in a QEA™ method. Two sequences producing two
fragments of identical end subsequences and length can be
10 discriminated by recognizing a third subsequence present in
one of the fragments but not in the other. In one
alternative, a labeled probe recognizing this third
subsequence can be added before detection to generate unique
signals from the fragment containing that subsequence. In
15 another alternative, a probe can be added before
amplification which prevents amplification of the fragment
with the third subsequence and which thereby removes
(suppresses) its signal. By way of example, such a probe can
be either an RE for recognizing and cutting the fragment with
20 the third subsequence, or a PNA, or modified DNA probe which
will hybridize with the third subsequence and prevent its PCR
amplification.

Further RE/ligase alternative methods increase
sample sequence discrimination in QEA™ experiments, for
25 example, by recognizing target subsequences longer or less
limited than those recognized by REs; such target
subsequences are termed herein effective target subsequences
or effective subsequences. This added information can often
discriminate two sample sequences producing fragments having
30 identical original end subsequences and lengths. The
effective subsequences are used in the computer implemented
database lookup methods of this invention in a manner similar
to the use of target subsequences. In one alternative,
termed herein a SEQ-QEA™ method, the target subsequences
35 recognized are effectively lengthened by using an
amplification primer with an internal Type IIS RE recognition
site so positioned that the Type IIS RE cuts the amplified

fragments in a manner producing a second overhang contiguous
with the recognition site of the initial RE. The sequence of
the second overhang concatenated with the initial target end
subsequence produces an effective subsequence that comprises,
5 and is longer than, the target subsequence recognized by the
RE. Alternatively, an effectively longer target subsequence
can be recognized by using phasing primers during PCR
amplification. The PCR amplification step can be divided
into several pools with each pool using one phasing
10 amplification primer constructed so as to recognize one or
more additional nucleotides beyond the original RE
recognition site. These additional nucleotides then
contribute to an effective subsequence that comprises the
target subsequence recognized by the RE.
15   In a SEQ-QEA™ method embodiment, an additional 4-8
bp subsequence is recognized at the end of a fragment by
digestion of a primer by a type IIS RE. This resulting
overhang is precisely contiguous with the RE cut end and is
sequenced in a standard manner, as by conventional Sanger
20 reactions. The additional subsequence information is
combined with the RE recognition subsequence to generate an
effective longer end target subsequence that is used as the
effective subsequence.
                The signals generated from the recognition
25 reactions of a QEA™ method experiment are analyzed by
computer methods. The analysis methods simulate a QEA™
method experiment using a database either of substantially
all known DNA sequences or of substantially all, or at least
a majority of, the DNA sequences likely to be present in a
30 sample to be analyzed and a description of the reactions to
be performed. The simulation results in a digest database
which contains for all possible signals that can be generated
the sample sequences responsible. Thereby, finding the
sequences that can generate a signal involves a look-up in
35 the simulated digest database. Computer implemented design
methods optimize the choice of target subsequences in the
QEA™ method reactions in order to maximize the information

produced in an experiment. For the tissue mode, the methods
maximize the number of sequences having unique signals by
which their quantitative presence can be unambiguously
determined. For the query mode, the methods maximize only
5 the number of sequences of interest having unique signals,
ignoring other sequences that might be present in a sample.

In QEA™ method embodiments wherein high stringency
hybridization is specified, such conditions generally
comprise a low salt concentration, equivalent to a
10 concentration of SSC (173.5 g. NaCl, 88.2 g. Na Citrate, H_2O
to 1 l.) of less than approximately 1 mM, and a temperature
near or above the T_m of the hybridizing DNA. In contrast,
conditions of low stringency generally comprise a high salt
concentration, equivalent to a concentration of SSC of
15 greater than approximately 150 mM, and a temperature below
the T_m of the hybridizing DNA.

In QEA™ method embodiments wherein DNA oligomers
are specified for performing functions, including
hybridization and chain elongation priming, alternatively
20 oligomers can be used that comprise those of the following
nucleotide mimics which perform similar functions.
Nucleotide mimics are subunits (other than classical
nucleotides) which can be polymerized to form molecules
capable of specific, Watson-Crick-like base pairing with DNA.
25 The oligomers can be DNA or RNA or chimeric mixtures or
derivatives or modified versions thereof. The oligomers can
be modified at the base moiety, sugar moiety, or phosphate
backbone. The oligomers may include other appending groups
such as peptides, hybridization-triggered cleavage agents
30 (see, e.g., Krol et al., 1988, BioTechniques 6:958-976), or
intercalating agents (see, e.g., Zon, 1988, Pharm. Res.
5:539-549). The oligomers may be conjugated to another
molecule, e.g., a peptide, hybridization triggered cross-
linking agent, transport agent, hybridization-triggered
35 cleavage agent, etc.

The oligomers may also comprise at least one
nucleotide mimic that is a modified base moiety which is

- 119 -

selected from the group including, but not limited to, 5-fluorouracil, 5-bromouracil, 5-chlorouracil, 5-iodouracil, hypoxanthine, xantine, 4-acetylcytosine, 5-(carboxyhydroxylmethyl) uracil, 5-carboxymethylaminomethyl-

5 2-thiouridine, 5-carboxymethylaminomethyluracil, dihydrouracil, beta-D-galactosylqueosine, inosine, N6-isopentenyladenine, 1-methylguanine, 1-methylinosine, 2,2-dimethylguanine, 2-methyladenine, 2-methylguanine, 3-methylcytosine, 5-methylcytosine, N6-adenine,

10 7-methylguanine, 5-methylaminomethyluracil, 5-methoxyaminomethyl-2-thiouracil, beta-D-mannosylqueosine, 5'-methoxycarboxymethyluracil, 5-methoxyuracil, 2-methylthio-N6-isopentenyladenine, uracil-5-oxyacetic acid (v), wybutoxosine, pseudouracil, queosine, 2-thiocytosine,

15 5-methyl-2-thiouracil, 2-thiouracil, 4-thiouracil, 5-methyluracil, uracil-5-oxyacetic acid methylester, 3-(3-amino-3-N-2-carboxypropyl) uracil, and 2,6-diaminopurine. The oligomers may comprise at least one modified sugar moiety selected from the group including, but

20 not limited to, arabinose, 2-fluoroarabinose, xylulose, and hexose. The oligomers may comprise at least one modified phosphate backbone selected from the group consisting of a phosphorothioate, a phosphorodithioate, a phosphoramidothioate, a phosphoramidate, a phosphordiamidate,

25 a methylphosphonate, an alkyl phosphotriester, and a formacetal or analog thereof.

The oligomer may be an α-anomeric oligomer. An α-anomeric oligomer forms specific double-stranded hybrids with complementary RNA in which, contrary to the usual β-units,

30 the strands run parallel to each other (Gautier et al., 1987, *Nucl. Acids Res.* 15:6625-6641).

Oligomers for use in a QEA™ method can be synthesized by standard methods known in the art, *e.g.*, by use of an automated DNA synthesizer (such as are commercially

35 available from Biosearch, Applied Biosystems, etc.). As examples, phosphorothioate oligonucleotides may be synthesized by the method of Stein et al. (1988, *Nucl. Acids*

Res. <u>16</u>:3209), methylphosphonate oligonucleotides can be prepared by use of controlled pore glass polymer supports (Sarin et al., 1988, *Proc. Natl. Acad. Sci. USA* <u>85</u>:7448-7451), etc.

5   In specific QEA™ method embodiments, it is preferable to use oligomers that can specifically hybridize to subsequences of a DNA sequence too short to achieve reliably specific recognition, such that a set of target subsequences is recognized. Further where PCR is used, since

10 *Taq* polymerase tolerates hybridization mismatches, PCR specificity is generally less than hybridization specificity. Where such oligomers recognizing short subsequences are preferable, they can be constructed in manners including, but not limited to, the following. To achieve reliable

15 hybridization to shorter DNA subsequences, degenerate sets of DNA oligomers can be used which are constructed of a total length sufficient to achieve specific hybridization with each member of the set containing a shorter sequence complementary to the common subsequence to be recognized. Alternatively, a

20 longer DNA oligomer can be constructed with a shorter sequence complementary to the subsequence to be recognized and with additional universal nucleotides or nucleotide mimics, which are capable of hybridizing to any naturally occurring nucleotide. Nucleotide mimics are sub-units which

25 can be polymerized to form molecules capable of specific, Watson-Crick-like base pairing with DNA. Alternatively, the oligomers may be constructed from DNA mimics which have improved hybridization energetics compared to naturally occurring nucleotides.

30   A preferred mimic is a peptido-nucleic acid ("PNA") based on a linked N-(2-aminoethyl)glycine backbone to which normal DNA bases have been attached (Egholm et al., 1993, *Nature* <u>365</u>:566-67). This PNA obeys specific Watson-Crick base pairing, but with greater free energy of binding and

35 correspondingly higher melting temperatures. Suitable oligomers may be constructed entirely from PNAs or from mixed PNA and DNA oligomers.

In QEA™ method embodiments wherein DNA fragments are separated by length, any length separation means known in the art can be used. One alternative separation means employs a sieving medium for separation by fragment length 5 coupled with a force for propelling the DNA fragments though the sieving medium. The sieving medium can be a polymer or gel, such a polyacrylamide or agarose in suitable concentrations to separate 10-1000 bp DNA fragments. In this case the propelling force is a voltage applied across the 10 medium. The gel can be disposed in electrophoretic configurations comprising thick or thin plates or capillaries. The gel can be non-denaturing or denaturing. Alternately, the sieving medium can be such as used for chromatographic separation, in which case a pressure is the 15 propelling force. Standard or high performance liquid chromatographic ("HPLC") length separation means may be used. An alternative separation means employs molecular characteristics such as charge, mass, or charge to mass ratio. Mass spectrographic means capable of separating 10- 20 1000 bp fragments may be used.

DNA fragment lengths determined by such a separation means represent the physical length in base pairs between target subsequences, after adjustment for biases or errors introduced by the separation means and length changes 25 due to experimental variables (e.g., presence of a detectable label, ligation to an adopter molecule). A represented length is the same as the physical length between occurrences of target subsequences in a sequence from said database when both said lengths are equal after applying corrections for 30 biases and errors in said separation means and corrections based on experimental variables. For example, represented lengths determined by electrophoresis can be adjusted for mobility biases due to average base composition or mobility changes due to an attached labeling moiety and/or adapter 35 strand by conventional software programs, such as Gene Scan Software from Applied Biosystems, Inc. (Foster City, CA).

In QEA™ method embodiments wherein DNA fragments
are labeled and detected, any compatible labeling and
detection means known in the art can be used. Advances in
fluorochromes, in optics, and in optical sensing now permit
5 multiply labeled DNA fragments to be distinguished even if
they completely overlap in space, as in a spot on a filter or
a band in a gel. Results of several recognition reactions or
hybridizations can be multiplexed in the same gel lane or
filter spot. Fluorochromes are available for DNA labeling
10 which permit distinguishing 6-8 separate products
simultaneously (Ju et al., 1995, *Proc. Natl. Acad Sci. USA*
92:4347-4351).

Exemplary fluorochromes adaptable to a QEA™ method
and methods of using such fluorochromes to label DNA are
15 described in § 6.1.12.4 (entitled "Fluorescent Labels For
QEA™ methods").

Single molecule detection by fluorescence (Eigen et
al., 1994, *Proc. Natl. Acad Sci. USA* 91:5740-5747) can also
be adapted for use.

20   In QEA™ method embodiments wherein intercalating
DNA dyes are utilized to detect DNA, any suitable dye known
in the art can be used. In particular such dyes include, but
are not limited to, ethidium bromide, propidium iodide,
Hoechst 33258, Hoechst 33342, acridine orange, and ethidium
25 bromide homodimers. Such dyes also include POPO, BOBO, YOYO,
and TOTO from Molecular Probes (Eugene, OR).

Alternative sensitive detection means available
include silver staining of polyacrylamide gels (Bassam et
al., 1991, *Analytic Biochemistry* 196:80-83), and the use of
30 intercalating dyes. In this embodiment, the gel can be
photographed and the photograph scanned by scanner devices
conventional in the computer art to produce a computer record
of the separated and detected fragments. A further
alternative is to blot an electrophoretic separating gel onto
35 a filter (*e.g.*, nitrocellulose) and then to apply any
visualization means known in the art to visualize adherent
DNA. See, *e.g.*, Kricka et al., 1995, Molecular Probing,

Blotting, and Sequencing, Academic Press, New York. In particular, visualization means utilizing secondary reactions with one or more reagents or enzymes can be used.

5 A preferred separation and detection apparatus for use in a QEA™ method is found in copending U.S. Patent Application Serial No. 08/438,231 filed May 9, 1995, which is herein incorporated by reference in its entirety. Other detection means adaptable to a QEA™ method include the commercial electrophoresis machines from Applied Biosystems 10 Inc. (Foster City, CA), Pharmacia (ALF), Hitachi, Licor. The Applied Biosystems machine is preferred among these as it is the only machine capable of simultaneous 4 dye resolution.

In the following subsections and the accompanying examples sections a QEA™ method embodiment is described in 15 detail.

5.4.2. DETAILS OF A QUANTITATIVE EXPRESSION ANALYSIS METHOD

This embodiment of a QEA™ method preferably 20 generates one or more signals unique to each cDNA sequence in a sample containing a mixture of cDNAs, and to quantitatively relate the strength of such a signal or signals to the relative amount of that cDNA sequence in the sample. Less preferably, the signals uniquely determine only sets of a 25 small number of sequences, typically 2-10 sequences. QEA™ method signals comprise an indication of the presence of pairs of target subsequences and the length between pairs of adjacent subsequences in a DNA sample. Signals are generated in a manner permitting straightforward automation with 30 existing laboratory robots. For simplicity of disclosure, and not by way of limitation, the detailed description of this method is directed to the analysis of samples comprising a plurality of cDNA sequences. It is equally applicable to samples comprising a single sequence or samples comprising 35 sequences of other types of DNA or nucleic acids generally.

While described in terms of cDNA hereinbelow, it will be understood that the DNA sample can be any DNA, e.g.,

cDNA and/or genomic DNA, and preferably comprises a mixture of DNA sequences. In specific embodiments, the DNA sample is an aliquot of cDNA of total cellular RNA or total cellular mRNA, most preferably derived from human tissue. The human
5 tissue can be diseased or normal.

The cDNA, or the mRNA from which it is synthesized, should be present at some threshold level in order to generate signals, this level being determined to some degree by the conditions of a particular QEA™ method experiment.
10 For example, such a threshold is that preferably at least 1000, and more preferably at least 10,000, mRNA molecules of the sequence to be detected be present in a sample. In the case where one or only a few mRNAs of a type of interest are present in each cell of a tissue from which it is desired to
15 derive the sample mRNA, at least a corresponding number of such cells should be present in the initial tissue sample. In a specific embodiment, the mRNA detected is present in a ratio to total sample RNA of $1:10^5$ to $1:10^3$. With a lower ratio, more molecular amplification can be performed during a
20 QEA™ method experiment.

The cDNA sequences occurring in a tissue derived pool include short untranslated sequences and translated protein coding sequences, which, in turn, may be a complete protein coding sequence or some initial portion of a coding
25 sequence, such as an expressed sequence tag. A coding sequence may represent an as yet unknown sequence or gene or an already known sequence or gene entered into a DNA sequence database. Exemplary sequence databases include those made available by the National Center for Biotechnology
30 Information ("NCBI") (Bethesda, MD) (GenBank) and by the European Bioinformatics Institute ("EMBL") (Hinxton Hall, UK).

A QEA™ method is also applicable to samples of genomic DNA in a manner similar to its application to cDNA.
35 In gDNA samples, information of interest includes occurrence and identity of translocations, gene amplifications, loss of heterozygosity for an allele, etc. This information is of

interest in cancer diagnosis and staging. In cancer
patients, amplified sequences might reflect an oncogene,
while loss of heterozygosity might reflect a tumor suppressor
gene. Such sequences of interest can be used to select
5 target subsequences and to predict signals generated by a
QEA™ method experiment. Even without prior knowledge of the
sequences of interest, detection and classification of the
QEA™ method signal patterns is useful for the comparison of
normal and diseased states or for observing the progression
10 of a disease state.

     Classification of QEA™ method signal patterns, in
an exemplary embodiment, can involve statistical analysis to
determine significant differences between patterns of
interest. This can involve first grouping samples that are
15 similar in one of more characteristics, such characteristics
including, for example, epidemiological history,
histopathological state, treatment history, etc. Signal
patterns from similar samples are then compared, e.g., by
finding the average and standard deviation of each individual
20 signals. Individual signal which are of limited variability,
e.g., for which the standard deviation is less than the
average, then represent genetic constants of samples of this
particular characteristic. Such limited variability signals
from one set of tissue samples can then be compared to
25 limited variability signals from another set of tissue
samples. Signals which significantly differ in this
comparison then represent significant differences in the
genetic expression between the tissue samples and are of
interest in reflecting the biological differences between the
30 samples, such as the differences caused by the progression of
a disease. For example, a significant difference in
expression is detected with the difference in the genetic
expression between two tissues exceed the sum of the standard
deviation of the expressions in the tissues. Other standard
35 statistical comparisons can also be used to establish level
of expression and the significance of differences in levels
of expressions.

Target subsequence choice is important in the
practice of a QEA™ method. The two primary considerations
for selecting subsequences are, first, redundancy, that is,
that there be enough subsequence pair hits per gene that a
5 unique signal is likely to be generated for each sample
sequence, and second, resolution, that is, that there not be
so many primer pairs hitting with very similar lengths in a
sample that the signals cannot be discriminated. For
sufficient redundancy, it is preferable that there be on
10 average, approximately three pair hits per gene or DNA
sequence in the sample. It is highly preferable that there
be at least one pair hit per each gene In a test of a
database of eukaryotic expressed sequences, it has been found
that an average value of three hits per gene appears to be
15 generally a sufficient guarantee of this minimum criterion.
Sufficient resolution depends on the separation and
detection means chosen. For a particular choice of
separation and detection means, a recognition reaction
preferably should not generate more fragments than can be
20 separated and distinguishably detected. In a preferred
embodiment, gel electrophoresis is the separation means used
to separate DNA fragments by length. Existing
electrophoretic techniques allow an effective resolution of
three base pair ("bp") length differences in sequences of up
25 to 1000 bp length. Given knowledge of fragment base
composition, effective resolution down to 1 bp is possible by
predicting and correcting for the small differences in
mobility due to differing base composition. However and
without limitation, an easily achievable three bp resolution
30 is assumed by way of example in the description of the QEA™
method. It is preferable for increased detection efficiency
that the distinguishably labeled products from as many
recognition reactions as possible be combined for separation
in one gel lane. This combination is limited by the number
35 of labels distinguishable by the employed detection means.
Any alternative means for separation and detection of DNA
fragments by length, preferably with resolution of three bp

or better, can be employed. For example, such separation means can be thick or thin plate or column electrophoresis, column chromatography or HPLC, or physical means such as mass spectroscopy.

5   The redundancy and resolution criteria are probabilistically expressed in Eqns. 1 and 2 in an approximation adequate to guide subsequence choice. In these equations the number of genes in the cDNA sequence mixture is N, the average gene length is L, the number of target

10 subsequence pairs is M (the number of pairs of recognition means), and the probability of each target subsequence hitting a typical gene is p. Since each target subsequences is preferably selected to independently hit each pooled sequence, the probability of an arbitrary subsequence pair

15 hitting is then p². Eqn. 1 expresses the redundancy condition of three hits per gene, assuming the probabilities of target subsequence hits are independent.

$$Mp^2 = 3 \qquad\qquad (1)$$

20 Eqn 2 expresses the resolution condition of having fragments with lengths no closer on average than 3 base pairs. This equation approximates the actual fragment length distribution with a uniform distribution.

25
$$\frac{L}{Np^2} = 3 \qquad\qquad (2)$$

Given expected values of N, the number of sequences in the library or pool to analyze (library complexity), and L, the average expressed sequence (or gene) length, Eqns 1 and 2 are

30 solved for the subsequence hit probability and number of subsequences required. This solution depends on the particular redundancy and resolution criteria dictated by the particular experimental method chosen to implement the QEA™ method. Alternative values may be required for other

35 implementations of a QEA™ method.

For example, it is estimated that the entire human genome contains approximately 10^5 protein coding sequences with an average length of 2000. The solution of Eqns. 1 and 2 for these parameters is p = 0.082 and M = 450. Thereby the

5 gene expression of all genes in all human tissues can be analyzed with a tissue mode QEA™ method using 450 target subsequence pairs, each subsequence having an independent probability of occurrence of 8.2%. In an embodiment in which eight fluorescently labeled subsequence pairs can be

10 optically distinguished and detected per electrophoresis lane, such as is possible when using the separation and detection apparatus described in copending U.S. Patent Application Serial No. 08/438,231 filed May 9, 1995, 450 reactions can be analyzed in only 57 lanes. Thereby only one

15 electrophoresis plate is needed in order to completely determine all human genome expression levels. Since the best commercial machines known to the applicants can discriminate only four fluorescent labels in one lane, a corresponding increase in the number of lanes is required to perform a

20 complete genome analysis with such machines.

As a further example, it is estimated that a typically complex human tissue expresses approximately 15,000 genes. The solution for N = 15000 and L = 2000 is p = 0.21 and M = 68. Thus expression in a typical tissue can be

25 analyzed with a tissue mode QEA™ method using 68 target subsequence pairs, each subsequence having an independent probability of occurrence of 21%. Assuming 4 subsequence pairs can be run per gel electrophoresis lane, the 68 reactions can be analyzed in 17 lanes in order to determine

30 the gene expression frequencies in any human tissue. Thus it is clear that this method leads to greatly simplified quantitative gene expression analysis within the capabilities of existing electrophoretic systems.

These equations provide an adequate guide to

35 picking subsequence pairs. Typically, preferred probabilities of target subsequence occurrence are from approximately 0.01 to 0.30. Probabilities of occurrence of

- 129 -

subsequences and RE recognition sites can be determined from databases of DNA sample sequences. Appropriate target subsequences can be selected from these tables. Computer implemented, QEA™ method experimental design methods can then
5 optimize this initial selection.

Another use of a QEA™ method is to compare directly the expression of only a few genes, typically 1 to 10, between two different tissues, the query mode, instead of seeking to determine the expression of all genes in a tissue,
10 the tissue mode. In this query mode, a few target subsequences are selected to identify the genes of interest both among themselves and from all other sequences possibly present. The computer design methods described hereinbelow can make this selection. If 4 subsequence pairs are
15 sufficient for identification, then the fragments from the 4 recognition reactions performed on each tissue are preferably separated and detected on two separate lanes in the same gel. If 2 subsequence pairs are sufficient for identification, the two tissues are preferably analyzed in the same gel lane.
20 Such comparison of signals from the same gel improves quantitative results by eliminating measurement variability due to differences between separate electrophoretic runs. For example, expression of a few target genes in diseased and normal tissue samples can be rapidly and reliably analyzed.
25   A query mode of a QEA™ method is also useful even if the sequences of the particular genes of interest are not yet known. For example, fluorescent traces produced by subjecting separate samples to gel electrophoretic separation means and then fluorescent detection means are compared to
30 identify feature differences. Such differentially expressed features created in a particular recognition reaction are then retrieved from the gel by methods known in the art (e.g. electro-elution from the gel) and their contained DNA fragments are analyzed by conventional techniques, such as by
35 sequencing. If partial, such sequences can then be used as probes (e.g., in PCR or Southern blot hybridization) to recover full-length sequences. In this manner, QEA™ method

techniques can guide the discovery of new differentially expressed cDNA or of changes of the state of gDNA. The sequences of the newly identified genes, once determined, can then be used to guide QEA™ method target subsequence choice

5 for further analysis of the differential expression of the new genes.

Two specific embodiments of a QEA™ method are described herein. The specific embodiments described herein use REs to recognize and cleave target subsequences in the

10 sample DNA. In one implementation, the desired doubly cut fragments are amplified by an amplification means in order to dilute remaining, unwanted singly cut fragments. Alternatively, the singly cut fragments are removed by physical means (e.g., hydroxyapatite column separation) or

15 enzymatic means (e.g., single strand specific nucleases). In another implementation, the unwanted singly cut ends are removed by a removal means from the desired doubly cut fragments without an amplification step, as described in § 5.4.3.2 (entitled "Second Alternative RE Embodiment"). For

20 these implementations, RE recognition sites define the possible target subsequences and are selected in a manner similar to the above in order to meet the previous probability or occurrence and independence criteria. The probabilities of occurrence of various RE recognition sites

25 are determined from a database of potential sample sequences, and those REs are chosen with recognition sequences whose probabilities of occurrence meet the criterion of Eqns. 1 and 2 as closely as possible. If multiple REs satisfy the selection criteria, a subset is selected by including only

30 those REs with independently occurring recognition sequences, determined, for example, by using conditional probabilities. Checking for independence can be done, by, for example, checking that the conditional probability for a hit by any selected pair of subsequences is the product of the

35 probabilities of the individual subsequence hit probabilities. An initial choice can be optionally optimized by the computer implemented experimental design methods.

- 131 -

A number, R_e, of REs are preferably selected so that the number of RE pairs is approximately M, where the relation between M and R_e is given by Eqn. 3.

5
$$M = \frac{R_e(R_e + 1)}{2} \qquad (3)$$

For example, a set a set of 20 acceptable REs results in 210 subsequence pairs.

10   There are numerous REs currently available whose recognition sequences have a wide range of occurrence probabilities, from which REs can be selected for the QEA™ method. A sample of these are presented in Example 6.1.12.3 (entitled "Preferred QEA™ Method Adapters And RE Pairs").

15   Restriction endonucleases ("RE") generally bind with specificity only to their short four to eight bp recognition sites, cleaving the DNA preferably with 4 bp complementary sequences. It is preferable that REs used in this embodiment produce overhangs characteristic of the particular RE. Thus REs, such as those known as class IIS
20   restriction enzymes, which produce overhangs of unknown sequence are less preferable. Class IIS REs are adaptable to generate short subsequences which may be sequenced to increase QEA™ method resolution by extending initial target subsequences into longer effective target subsequences. This
25   alternative embodiment is known as the SEQ-QEA™ method (see Section 5.4.4). Phasing primers can also be used to recognize longer effective target subsequences. Further, ligases, which are used in a QEA™ method to ligate an adapter strand to a digested terminus, are highly specific in their
30   hybridization requirements; even one bp mismatch near the ligation site will prevent ligation (U.S. Patent 5,366,877, Nov 22, 1994, to Keith et al.; U.S. Patent 5,093,245, Mar 3, 1992, to Keith et al.).

35   QEA™ method experiments are also adaptable to distinguish sequences into small sets, typically comprising 2 to 10 sequences, which require fewer target subsequence

pairs. Such coarser grain analysis of gene expression or genomic composition requires fewer recognition reactions and analysis time. Alternatively, smaller numbers of target subsequence pairs can be optimally chosen to distinguish
5 individually a specific set of genes of interest from all the other genes in the sample. These target subsequences can be chosen either from REs that produce fragments from the desired genes.

   Detailed descriptions of exemplary implementations
10 for practicing QEA™ method recognition reactions and related computer implemented experimental analysis and design methods are presented in the following subsections followed by detailed experimental protocols in the Examples subsections. The implementations are illustrative and not limiting, as a
15 QEA™ method can be practiced by any method generating the previously described QEA™ method signals.

5.4.3. RE EMBODIMENTS OF A QEA™ METHOD
   Restriction endonuclease ("RE") embodiments of a
20 QEA™ method use implementations of simultaneous RE and ligase enzymatic reactions for generating labeled fragments of the genes or sequences to be analyzed. These fragments are then separated by length by a separation means and detected by a detection means to yield QEA™ method signals comprising the
25 identity of the REs cutting each fragment together with each fragment's length. The recognition reactions can specifically and reproducibly generate QEA™ method signals with good signal to noise ratios and without any intermediate extractions or buffer exchanges, which would hinder automatic
30 execution.

   REs bind with specificity to short DNA target subsequences, usually 4 to 8 bp long, that are termed recognition sites and are characteristic of each RE. REs that are used cut the sequence at (or near) these recognition
35 sites preferably producing characteristic ("sticky") ends with single-stranded overhangs, which usually incorporate part of the recognition site.

Preferred REs have a 6 bp recognition site and generate a 4 bp 5' overhang. If more fragments are desired for the analysis of a particular sample, REs with shorter recognition sites can be used, for example with 4 or 5 bp
5 recognition sites. The RE embodiments are also adaptable to a 2 bp 5' overhang, which is less preferred since 2 bp overhangs have a lower ligase substrate activity than 4 bp overhangs. All RE embodiments can be adapted to 3' overhangs of two and four bp. Further preferred REs have the following
10 additional properties. Their recognition sites and overhang sequences are preferably such that an adapter can be designed whose ligation does not recreate the recognition site. They preferably have sufficient activity below 37°C and are heat inactivated at 65°C. Heat inactivation is preferable so that
15 RE inactivation can be performed prior to adding PCR reagents and conducting the PCR reaction in the same vial. They preferably have low non-specific cutting and nuclease activities and cut to completion. Of course, REs selected for a particular experiment preferably have recognition sites
20 meeting the previously described occurrence and independence criteria. Preferred pair of REs for analyzing human and mouse cDNA are listed in § 6.1.12.3 (entitled "Preferred QEA™ Method Adapters and RE Pairs").

Only doubly cut sequence fragments are of interest,
25 and thus in all RE QEA™ method embodiments the desired doubly cut fragments are distinguished from the unwanted singly cut fragments. Singly cut fragments have a non-specific and non-reproducible length distribution derived from the distribution of overall cDNA lengths, which depends strongly
30 on cDNA synthesis conditions. Only the doubly cut fragments have a specific and reproducible length distribution dependent only on the DNA sequence analyzed and independent of cDNA synthesis conditions. To make this distinction, the preferred RE embodiment of a QEA™ method exponentially
35 amplifies doubly cut fragments, so that their signals quickly overwhelm signals from singly cut fragments, which are at

most linearly amplified.   PCR is the preferred amplification means.

Alternative amplification means known in the art are adaptable to a QEA™ method. If a removal means for
5 singly cut ends is not utilized in an embodiment, alternative amplification means should preferentially amplify doubly cut ends over singly cut ends in order that signals from singly cut ends be relatively suppressed. On the other hand, if a removal means for singly cut ends is utilized in an
10 embodiment, then alternative amplification means need have no amplification preference, as no singly cut ends are present at the amplification step. Known alternative amplification means are listed in Kricka et al., 1995, Molecular Probing, Blotting, and Sequencing, chap. 1 and table IX, Academic
15 Press, New York. Of these alternative means, those employing the T7 RNA polymerase are preferred.

Two other specific embodiments use a physical removal means to directly remove singly cut fragments, preferably before amplification. This can be accomplished,
20 e.g., by labeling DNA termini with a capture moiety prior to digestion. After digestion, the singly cut fragments are removed by contacting the sample with a binding partner of the capture moiety, affixed to a solid phase. The preferred removal means is biotin-streptavidin. Other removal means
25 adaptable to this embodiment of a QEA™ method include various haptens, which are removed by their corresponding antibodies. Exemplary haptens include digoxigenin, DNP, and fluorescein (Holtke et al., 1992, Sensitive chemiluminescent detection of digoxigenin labeled nucleic acids: a fast and simple
30 protocol for applications, Biotechniques 12(1):104-113 and Olesen et al., 1993, Chemiluminescent DNA sequencing with multiple labeling, Biotechniques 15(3):480-485). Alternatively, single stranded fragments can be removed by single stand specific column separation or single strand
35 specific nucleases.

RE embodiments of a QEA™ method use recognition moieties which are specifically ligated to RE cut sticky ends

- 135 -

so that in any one recognition reaction ends cut by a
particular RE receive a unique moiety. Recognition moieties
comprise oligomers capable of specifically hybridizing to the
RE generated sticky ends. In the preferred RE embodiment,
5 which uses PCR amplification, the recognition moieties also
provide primer means for the PCR.

     The recognition moieties also provide for labeling
and recognition of RE cut ends. For example, using a pair of
REs in one recognition reaction generates doubly cut
10 fragments some with the recognition sequence of the first RE
on both ends, some with the recognition sequence of the
second RE on both ends, and the remainder with one
recognition sequence of each RE on either end. Using more
REs generates doubly cut fragments with all pairwise
15 combinations of RE cut ends from adjacent RE recognition
sites along the sample sequences. All these cutting
combinations need preferably to be distinguished, since each
provides unique information on the presence of different
subsequences pairs present in the original DNA sequence.
20 Thus the recognition moieties preferably have unique labels
which label specifically each RE cut made in a reaction. As
many REs can be used in a single reaction as labeled
recognition moieties are available to uniquely label each RE
cut. If the detectable labeling in a particular system is,
25 for example, by fluorochromes, then fragments cut with one RE
have a single fluorescent signal from the one fluorochrome
associated with that RE, while fragments cut with two REs
have mixed signals, one from the fluorochrome associated with
each RE. Thus all possible pairs of fluorochrome labels are
30 preferably distinguishable. Alternatively, if certain target
subsequence information is not needed, the recognition
moieties need not be distinctively labeled. In embodiments
using PCR amplification, corresponding primers would not be
labeled.
35   If silver staining is used to recognize fragments
separated on an electrophoresis gel, no recognition moiety
need be labeled, as fragments cut by the various RE

combinations are not distinguishable. In this case, when PCR amplification is used, only primers are required.

The recognition reaction conditions are preferably selected, as described in § 6.1.12.1 (entitled "QEA™

5 Preferred RE Method"), so that RE cutting and recognition moiety ligation go to full completion: all recognition sites of all REs in the reaction are cut and ligated to a recognition moiety. It is more preferable, in general, to perform the recognition reactions according to Sec.

10 6.1.12.2.1 ("QEA™ Method Preferred For Use In A SEQ-QEA™ Method"). This more preferred protocol describes performing the RE/ligase and PCR reactions in a single reaction vessel, with at least one primer having a conjugated capture moiety, followed by cleanup of certain reaction products. In this

15 manner, the fragments generated from a sequence analyzed lie only between adjacent recognition sites of any RE in that reaction. No fragments remain which include any RE recognition site, since such a site is cut. Multiple REs can be used in one recognition reaction. Too many REs in one

20 reaction may cut the sequences too frequently, generating a compressed length distribution with many short fragments of lengths between 10 and a few hundred base pairs long. Such a distribution may not be resolvable by the separation means, for example gel electrophoresis, if the fragments are too

25 close in length, for example less than 3 bp apart on the average. Too many REs also may generate fragments of the same length and end subsequences from different sample sequences, thereby leading to non-unique signals. Finally, where fragment labels are to be distinguished, no more REs

30 can be used than can have distinguishably labeled sticky ends. These considerations limit the number of REs optimally useable in one recognition reaction. Preferably two REs are used, with one, three and four REs less preferable. Preferable pairs of REs for the analysis of human cDNA

35 samples are listed in § 6.1.12.3 (entitled "Preferred QEA™ Method Adapters and RE Pairs").

An additional level of signal specificity is
possible by selecting or suppressing fragments having a third
internal target subsequence. Additional information on the
presence or absence of specific internal subsequences can be
5 used along with the two end subsequences and the length
information to further distinguish between otherwise
identically classified fragments.

Other methods of providing third subsequence
information are described below which label or suppress
10 fragments with third subsequences. To select fragments with
a third internal subsequence, probes with distinguishable
labels which bind to this target subsequence are added to the
fragments prior to detection, and alternatively prior to
separation and detection. On detection, fragments with this
15 third subsequence present will generate a signal, preferably
fluorescent, from the probe. Such a probe can be a labeled
PNA or DNA oligomer. Short DNA oligomers may need to be
extended with a universal nucleotide or degenerate sets of
natural nucleotides in order to provide for specific
20 hybridization.

Fragments with a third subsequence can be
suppressed in various manners in embodiments using PCR
amplification. First, a probe hybridizing with this third
subsequence which prevents polymerase elongation in PCR can
25 be added prior to amplification. Then sequences with this
subsequence will be at most linearly amplified and their
signal thereby suppressed. Such a probe could be a PNA or
modified DNA oligomer (with the last nucleotide being a
ddNTP). Second, if the third subsequence is recognized by an
30 RE, this RE can be added to the RE-ligase reaction without
any corresponding specific primer. Fragments with the third
subsequence will be at most linearly amplified.

Both these alternatives can be extended to multiple
internal sequences by using multiple probes to recognize the
35 sequences or to disrupt exponential PCR amplification.

Construction of the recognition moieties, also
herein called adapters or linker-primers, is important and is

- 138 -

described here in advance of further details of the
individual recognition reaction steps. In the preferred
embodiment, the adapters are partially double stranded DNA
("dsDNA"). Alternatively, the adapters can be constructed as
5 oligomers of any nucleic acid, with corresponding properties
to the preferred DNA polymers. In an embodiment employing an
alternative amplification means, any polymer that can serve
with a template as a primer for that amplification means can
be used in that embodiment.

10   Figure 10A illustrates the DNA molecules involved
in the ligation reaction as conventionally indicated with the
5' ends of the top strands and the 3' ends of the bottom
strands at left. dsDNA 201 is a fragment of a sample cDNA
sequence with an RE cut at the left end generating,
15 preferably, a four bp 5' overhang 202. Adapter dsDNA 209 is
a synthetic substrate provided by a QEA™ method.

          The precise characteristics of adapter 209 are
selected in order to ensure that RE digestion and adapter
ligation preferably go to completion, that generation of
20 unwanted products and amplification biases are minimized, and
that unique labels are attached to cut ends (if needed).
Adapter 209 comprises strand 203, called a primer, and a
partially complementary strand 205, called a linker. The
primer is also known as the longer strand of the adapter, and
25 the linker is also known as the shorter strand of the
adapter.

          The linker, or shorter strand, links the end of a
cDNA cut by an RE to the primer, or longer strand, by
hybridization to the sticky overhang of the cut end and to
30 the primer in order that the primer can be ligated to dsDNA
201. Therefore, linker 205 comprises sequence 206
complementary to the sticky RE overhang 202 and sequence 207
complementary to the 3' end of primer 203. Sequence 206 is
preferably of the same length as the RE overhang. Sequence
35 207 is most preferably eight nucleotides long, less
preferably from 4 to 12 nucleotides long, but can be of any
length as long as the linker reliably hybridizes with only

one top primer in any one recognition reaction and has an
appropriate T_m (preferably less than approximately 68°C).
Linker 205 also preferably has no 5' terminal phosphate so
that it will not ligate to the bottom strand of dsDNA 201.

5 Lack of terminal phosphate also prevents the annealed
adapters from ligating to each other, forming dimers, and
thereby competing with adapter ligation to RE cut sample
fragments. Adapter dimers would also be amplified in a
subsequent amplification step generating unwanted fragments.

10 Terminal phosphates can be removed using phosphatases (e.g.,
alkaline phosphatase) known in the art, followed by
separation of the enzyme.

Further, the linker, or shorter strand, T_m should
preferably be less than primer 203 self-annealing T_m. This

15 ensures that subsequent PCR amplification conditions can be
controlled so that linkers present in the reaction mixture
will not hybridize and act as PCR primers, and, thereby,
generate spurious fragment lengths. The preferable T_m is less
than approximately 68°C.

20   Primer, or longer strand, 203 further has a 3' end
sequence 204 complementary to 3' end sequence 207 of bottom
linker 205. In a preferred aspect, in order that all RE cuts
are properly ligated to a unique top primer, in any single
reaction, each primer should be complementary to and

25 hybridize with only one linker 205. Consequently, all the
linkers in any one reaction mixture preferably have unique
sequences 207 for hybridizing with unique primers. In order
that the ligation reaction go to completion, primer 203
preferably should not recreate the recognition sequence of

30 any RE in the reaction mixture when it is ligated with cDNA
end 202. Primer 203 has no 5' terminal phosphate in order to
prevent any self-ligations. To minimize amplification of
undesired sequences, termed amplification noise, in any
subsequent PCR step it is preferred that primer 203 not

35 hybridize with any sequence present in the original sample
mixture. The T_m of primer 203 is preferably high, in the
range from 50° to 80°C, and more preferably above 68°C. This

- 140 -

ensures that the subsequent PCR amplification can be
controlled so that only primers and not linkers initiate new
chains. For example, this T_m can be achieved by use of a
primer having a combination of a G+C content preferably from
5 40-60%, most preferably from 55-60%, and a primer length most
preferably 24 nucleotides, and preferably from 18 to 30
nucleotides. Primer 203 is optionally labeled with
fluorochrome 208, although any DNA labeling system that
preferably allows multiple labels to be simultaneously
10 distinguished is usable in the QEA™ method.

　　　　Generally, the primer, or longer strand, are
constructed so that, preferably, they are highly specific,
free of dimers and hairpins, and form stable duplexes under
the conditions specified, in particular the desired T_m.
15 Software packages are available for primer construction
according to these principles, an example being OLIGO™
Version 4.0 For Macintosh from National Biosciences, Inc.
(Plymouth, MN). In particular, a formula for T_m can be found
in the OLIGO™ Reference Manual at Eqn. I, page 2.
20　　　　Figure 10B illustrates two exemplary adapters and
their component primers and linkers constructed according to
the above description. Adapter 250 is specific for the RE
BamHI, as it has a 3' end complementary to the 5' overhang
generated by BamHI. Adapter 251 is similarly specific for
25 the RE HindIII.

　　　　Example 6.1.12.3 (entitled "Preferred QEA™ Method
Adapters And RE Pairs") contains a more comprehensive, non-
limiting list of adapters that can be used according to the
QEA™ method. All synthetic oligonucleotides used in the QEA™
30 method are preferably as short as possible for their
functional roles in order to minimize synthesis costs.

　　　　Alternatively, adapters can be constructed from
hybrid primers which are designed to facilitate the direct
sequencing of a fragment or the direct generation of RNA
35 probes for *in situ* hybridization with the tissue of origin of
the DNA sample analyzed. Hybrid primers for direct
sequencing are constructed by ligating onto the 5' end of

existing primers the M13-21 primer, the M13 reverse primer,
or equivalent sequences. Fragments generated with such
hybrid adapters can be removed from the separation means and
amplified and sequenced with conventional systems. Such
5 sequence information can be used both for a previously known
sequence to confirm the sequence determination and for a
previously unknown sequence to isolate the putative new gene.
Hybrid primers for direct generation of RNA hybridization
probes are constructed by ligating onto the 5' end of
10 existing primers the phage T7 promoter. Fragments generated
with such hybrid adapters can be removed using the separation
means and transcribed into anti-sense RNA with conventional
systems. Such probes can be used for *in situ* hybridization
with the tissue of origin of the DNA sample to determine in
15 precisely what cell types a signal of interest is expressed.

A further alternative illustrated in Figure 10C is
to construct an adapter by self hybridization of single
stranded DNA in hairpin loop configuration 212. The
subsequences of loop 212 would have similar properties to the
20 corresponding subsequences of linker 205 and primer 203.
Exemplary hairpin loop 211 sequences are C_4 to C_{10}.

REs generating 3' overhangs are less preferred and
require the different adapter structure illustrated in Figure
11A. dsDNA 301 is a fragment of a sample cDNA cut with a RE
25 generating 3' sticky overhang 302. Adapter 309 comprises
primer, or longer strand, 304 and linker, or shorter strand,
305. Primer, or longer strand, 304 includes segment 306
complementary to and of the same length as 3' overhang 302
and section 307 complementary to linker 305. It also
30 optionally has label 308 which distinctively labels primer
304. As in the case of adapters for 5' overhangs, primer 304
has no 5' terminal phosphate, in order to prevent self-
ligations, and is such that no recognition site for any RE in
one recognition reaction is created upon ligation of the
35 primer with dsDNA 301. These condition ensure that the RE
digestion and ligation reactions go to completion. Primer
304 should preferably not hybridize with any sequence in the

- 142 -

initial sample mixture.   The T_m of primer 304 is preferably high, in the range from 50° to 80°C, and more preferably above 68°C.   This ensures the subsequent PCR amplification can be controlled so that only primers and not linkers

5 initiate new chains.   For example, this T_m can be achieved by using a primer having a G+C content preferably from 40-60%, most preferably from 55-60%, and a primer length most preferably of 24 nucleotide and less preferably of 18-30 nucleotides.   Each primer 304 in a reaction can optionally

10 have a distinguishable label 308, which is preferably a fluorochrome.

Linker, or shorter strand, 305 is complementary to and hybridizes with section 307 of primer 304 such that it is adjacent to 3' overhang 302.   Linker 305 is most preferably 8

15 nucleotides long, less preferably from 4-16 nucleotides, and has no terminal phosphates to prevent any self-ligation. This linker serves only to promote ligation specificity and reaction speed.   It does not perform the function of linking primer 304 to the cut dsDNA, as it did in the 5' case.

20 Further, linker 305 T_m should preferably be less than primer 304 self-annealing T_m.   This insures that subsequent PCR amplification conditions can be controlled so that linkers present in the reaction mixture will not hybridize and act as PCR primers, and, thereby, generate spurious fragment

25 lengths.

Figure 11B illustrates an exemplary adapter with its primer and linker for the case of the RE NlaIII.   As in the 5' overhang case, a 3' adapter can also be constructed from a hairpin loop configuration.

30 REs generating 5' and 3' overhangs are preferably not used in the same recognition reaction.   This is in order that a complementary primer hybridization site can be presented on each of the two strands of the product of the RE/ligase recognition reaction.

35 Turning now to a detailed description of a preferred RE embodiment of the QEA™ method recognition reactions, the steps of this preferred embodiment comprise,

- 143 -

first, simultaneously cleaving a mixed DNA sample (e.g., one
of the populations of proteins being assayed for interaction
by the method of the invention with another protein
population, or a pooled group of cDNAs encoding interacting
5 proteins identified in the assay) with one or more REs and
ligating recognition moieties on the cut ends, second,
amplifying the twice cut fragments, if necessary, and third,
separating the fragments by length and detecting the lengths
and labels, and the identities of the REs cutting each
10 fragment. Following the amplification step, optional steps
to remove unwanted singly stranded DNA fragments prior to
detection can increase the signal to noise ratio of the
following detection. Two alternative RE embodiments are
described in following subsections. The number of REs and
15 associated adapters preferably are limited so that both a
compressed length distribution consisting of shorter
fragments is avoided and enough distinguishable labels are
available for all the REs used. Alternatively, REs can be
used without associated adapters in order that the amplified
20 fragments not have the associated recognition sequences.
Absence of these sequences can be used to additionally
differentiate genes that happen to produce fragments of
identical length with particular REs.

        A cDNA sample is prepared prior to carrying out a
25 QEA™ method by removal of terminal phosphates from all the
cDNA. This is important to improve the signal to noise ratio
in the subsequent fragment length separation and detection by
eliminating amplification of unwanted, singly cut fragments.
Significant background signals arise from exponential
30 amplification of singly cut fragments whose blunt ends have
ligated to form a single dsDNA with two cut ends, an
apparently doubly cut fragment, which is exponentially
amplified like a normal doubly cut fragment. Since cDNA
lengths vary depending on synthesis condition, these
35 unwanted, apparently doubly cut fragments have a wide range
of lengths and produce a diffuse background on gel
electrophoresis which obscures sharp bands from the normally

doubly cut fragments. This background can be eliminated by
preventing blunt end ligation of singly cut fragments by
initially removing all terminal phosphates from the cDNA
sample, without otherwise disrupting the integrity of the
5 cDNA.

　　　　Terminal phosphate removal is preferably done with
a phosphatase. To prevent interference with the intended
ligation of adapters to doubly cut fragments, the phosphatase
activity preferably is removed prior to the RE digestion and
10 adapter ligation step. To avoid any phosphatase separation
or extraction step, the preferred phosphatase is a heat
labile alkaline phosphatase which is heat inactivated prior
to the RE/ligase step. A preferred phosphatase comes from
cold living Barents Sea (arctic) shrimp (U.S. Biochemical
15 Corp.) ("shrimp alkaline phosphatase" or "SAP"). Terminal
phosphate removal need be done only once for each population
of cDNA being analyzed.

　　　　In other embodiments additional phosphatases may be
used for terminal phosphate removal, such as calf intestinal
20 phosphatase-alkaline from Boehringer Mannheim (Indianapolis,
IN). Those that are not heat inactivated require the
addition of a step to separate the phosphatase from the cDNA
before the recognition reactions, such as by phenol-
chloroform extraction.
25　　　　Preferably, the prepared cDNA is then separated
into batches of from 1 picogram ("pg") to 200 nanograms
("ng") of cDNA each, and each batch is separately processed
by the further steps of the method. For a tissue mode
experiment, to analyze gene expression, preferably from a
30 majority of expressed genes, from a single human tissue
requires determination of the presence of about 15,000
distinct cDNA sequences. By way of example, one sample is
divided into approximately 50 batches, each batch is then
subject to the RE/ligase recognition reaction and generates
35 approximately 200-500 fragments, and more preferably 250 to
350 fragments of 10 to 1000 bp in length, the majority of
fragments preferably having a distinct length and being

uniquely derived from one cDNA sequence. A preferable
example analysis would entail 50 batches generating
approximately 300 bands each.

      For the query mode, fewer recognition reactions are
5 employed since only a subset of the expressed genes are of
interest, perhaps approximately from 1 to 100. The number of
recognition reactions in an experiment may then number
approximately from 1 to 10 and an appropriate number of cDNA
batches is prepared.

10   Following cDNA preparation, the next step is
simultaneous RE cutting of and adapter ligation to the sample
cDNA sequences. The prepared sample is cut with one or more
REs. The amount of RE enzyme in the reaction is preferably
approximately a 10 fold unit excess. Substantially greater
15 quantities are less preferred because they can lead to star
activity (non-specific cutting) while substantially lower
quantities are less preferred because they will result in
less rapid and only partial digestion, and hence incomplete
and inaccurate characterization of the subsequence
20 distribution.

      In the same reaction, adapters and ligase enzyme
are present for simultaneous adapter ligation to the RE cut
ends. The method is adaptable to any ligase that is active
in the temperature range 10 to 37°C. T4 DNA ligase is the
25 preferred ligase. In other embodiments, cloned T4 DNA ligase
or T4 RNA ligase can also be used. In a further embodiment,
thermostable ligases can be used, such as Ampligase™
Thermostable DNA Ligase from Epicenpre (Madison, WI), which
has a low blunt end ligation activity. These ligases in
30 conjunction with the repetitive cycling of the basic thermal
profile for the RE-ligase reaction, described in the
following, permit more complete RE cutting and adapter
ligation.

      Ligase activity can both generate unwanted products
35 and also, if an RE recognition site is regenerated, can cause
an endless cycle of further cutting and ligation. Terminal
phosphate removal during cDNA preparation prevents spurious

- 146 -

ligation of the blunt other ends of singly cut cDNA (and
subsequent exponential amplification of the results). Other
unwanted products are fragment concatamers formed when the
sticky ends of cut cDNA fragments hybridize and ligate. Such
5 fragment concatamers are removed by keeping the restriction
enzymes active during ligation, thus cutting unwanted
concatamers once they form. Further, adapters, once ligated,
terminate further RE cutting, since adapters are selected
such that RE recognition sites are not recreated. A high
10 molar excess of adapters also is preferable since it limits
concatamer formation by driving the RE and ligase reactions
toward complete digestion and adapter ligation. Finally,
unwanted adapter self-ligation is prevented since primers and
linker also lack terminal phosphates (preferably due to
15 synthesis without phosphates or less preferably due to
pretreatment thereof with phosphatases).

The temperature profile of the RE/ligase reaction
is important for achieving complete cutting and ligation.
The preferred protocol has several stages. The first stage
20 is at the optimum RE temperature to achieve substantially
complete cutting, for example 37°C for 30 minutes. The
second stage is a ramp at -1°C/min down to a third stage
temperature for substantially compete annealing of adapters
to the sticky cut ends and primer ligation. During this
25 ramp, cutting and ligation continue. The third stage is at
the optimum temperature for adapter annealing and ligation to
the sticky ends, and is, for example, at 16°C for 60 minutes.
The fourth stage is again at the optimum RE to achieve
complete cutting of all recognition sites, for example at
30 37°C for 15 minutes. The fifth stage is to heat inactivate
the ligase and, preferably, also the RE enzymes, and is, for
example, 10 minutes at or above 65°C. If the PCR reaction is
not to be immediately performed, the results are held at 4°C.
If the PCR amplification is to be immediately performed, as
35 in the preferred single tube protocol of Sec. 6.1.12.2.1
("QEA™ Method Preferred For Use In A SEQ-QEA™ Method"), this
fifth stage is at 72°C for 20 minutes.

A less preferred profile involves repetitive
cycling of the first four stages of the temperature protocol
described above, that is from an optimum RE temperature to
optimum annealing and ligation temperatures, and back to an
5 optimum RE temperature. The additional cycles further drive
the RE/ligase reactions to completion. In this embodiment,
it is preferred to use thermostable ligase enzymes. The
majority of restriction enzymes are active at the
conventional 16°C ligation temperature and hence prevent
10 unwanted ligation events without thermal cycling. However,
temperature profiles consisting of optimum ligation
conditions interspersed with optimum RE cutting conditions
cause both enzymatic reactions to proceed more rapidly than
one constant temperature. An exemplary profile comprises
15 periodically cycling between a 37°C optimum RE temperature to
a 16°C optimum annealing and ligation temperature at a ramp
of ~1°C/min, and then back to the 37°C optimum RE
temperature. Following completion of approximately 2 to 4 of
these temperature cycles, the RE and ligase enzymes are heat
20 inactivated by a final stage at 65°C for 10 minutes. This
avoids the need for separation or extractions between steps.
The results are held at 4°C.

These thermal profiles are easily controlled and
automated by the use of commercially available computer
25 controlled thermocyclers, for example from MJ Research
(Watertown, MA) or Perkin Elmer (Norwalk, CT).

These reaction conditions are designed to achieve
substantially complete cutting of all RE recognition sites
present in the analyzed sequence mixture and complete
30 ligation of reaction terminating adapters on the cut ends,
each adapter being unique in one reaction for a particular RE
cut end. The fragments generated are limited by adjacent RE
recognition sites and no fragment includes internal
undigested sites. Further, a minimum of unwanted self-
35 ligation products and concatamers is formed.

Following the RE/ligase step is amplification of
the doubly cut cDNA fragments. Although PCR protocols are

described in the exemplary embodiment, any amplification
method that selects fragments to be amplified based on end
sequences is adaptable to a QEA™ method (see above). With
high enough sensitivity of detection means, or even single
5 molecule detection means, the amplification step can be
dispensed with entirely. This is preferable as amplification
inevitably distorts the quantitative response of the method.

The PCR amplification protocol is designed to have
maximum specificity and reproducibility. First, the PCR
10 amplification produces fewer unwanted products if the
amplification steps occur at a temperature above the T_m of the
shorter linker so that it cannot initiate unwanted DNA
strands. The linker is preferably melted by an initial
incubation at 72°C without the Taq polymerase enzyme or dNTP
15 substrates present. A further incubation at 72°C for 10
minutes with Taq polymerase and dNTPs is performed in order
to complete partial double strands to complete double
strands. Alternatively, linker melting and double strand
completion can be performed by a single incubation at 72°C
20 for 10 minutes with Taq polymerase. Subsequent PCR
amplification steps are carried out at temperatures
sufficiently high to prevent re-hybridization of the bottom
linker.

Second, primer strand 203 of Figure 10A (and 304 of
25 Figure 11A) are typically used as PCR primers. They are
preferably designed for high amplification specificity and
not to hybridize with any native cDNA species to be analyzed.
They have high melting temperatures, preferably above 50°C
and most preferably above 68°C, to ensure specific
30 hybridization with a minimum of mismatches.

Third, the protocol's temperature profile is
preferably designed for specificity and reproducibility. A
preferred profile is 95°C for 30 seconds, then 57°C for 1
minute, and then 72°C for 2 minutes. High annealing
35 temperatures minimize primer mis-hybridizations. Longer
extension times reduce PCR bias in favor of smaller
fragments. Longer melting times reduces PCR amplification

bias in favor of high G+C content. Further, large
amplification volumes are preferred to reduce bias.
Sufficient amplification cycles are performed, typically
between 15 and 30 cycles.

5   Any other techniques designed to raise specificity,
yield, or reproducibility of amplification are applicable to
this method. One preferred technique is to include Betaine
(Sigma) in both the RE/ligase reaction and in the PCR
amplification. Another technique that can be used is the use
10 of 7-deaza-2'-dGTP in the PCR reaction in place of dGTP.
This has been shown to increase PCR efficiency for G+C rich
targets (Mutter et al., 1995, *Nuc. Acid Res.* 23:1411-1418).
As a further example, another technique that can be used is
the addition of tetramethylammonium chloride to the reaction
15 mixture, which has the effect of raising the T_m (Chevet et
al., 1995, *Nucleic Acids Research* 23(16):3343-3344).

           In a particular method of performing the PCR
amplification, each RE/ligase reaction sample is sub-divided
into multiple aliquots, and each aliquot is amplified with a
20 different number of cycles. Multiple amplifications with an
increasing number of amplification cycles, for example 10,
15, and 20 cycles, are preferable. Amplifications with a
lower number of cycles detect more prevalent messages in a
more quantitative manner. Amplification with a higher number
25 of cycles detect the presence of less prevalent genes but
less quantitatively. Multiple amplifications also serve as
controls for checking the reliability and quantitative
response of the process by comparing the size of the same
signal in each amplification.

30   Other methods of performing the PCR amplification
are more suited to automation. For example, the content of a
reaction vial can be configured as follows. First, 40 μl of
the PCR mix without Mg ions is added followed by a wax bead
that melts approximately at 72°C, such as Ampliwax beads
35 (Perkin-Elmer, Norwalk, CT). This bead is melted at 75°C for
5 minutes and solidified at 25°C for 10 minutes. A preferred
wax is a 90:10 mixture of Paraffin:Chillout™ 14. The

- 150 -

paraffin is a highly purified paraffin wax melting between 58
°C and 60 °C such as can be obtained from Fluka Chemical,
Inc. (Ronkonkoma, N.Y.) as Paraffin Wax cat. no. 76243.
Chillout™ 14 Liquid Wax is a low melting, purified paraffin
5 oil available from MJ Research. It is preferred to coat the
upper sides of the reaction tubes with this solidified wax,
carefully add the PCR mix, then melt this wax onto the PCR
mix by the temperature protocol in Sec. 6.1.12.2.1, which
beginning with a 2 min incubation at 72 °C then decreases the
10 temperature by 5 °C every 2 min until 25 °C is reached.
Then, the RE/ligase mix with Mg ions is added. The RE/ligase
and PCR reactions are carried out by following the preferred
temperature profile in Figure 22D. In this arrangement in
the same vial, the RE/ligase reactions can first be
15 performed. The incubation at 72°C for 20 minutes permits the
wax layer separating the mixtures to melt, allows the
RE/ligase mixture to mix with the PCR mix, and allows
completion of the partial double strands to complete double
strands. Then sufficient PCR cycles are performed, typically
20 between 15 and 30 cycles. This single tube implementation is
well adapted to automation. Other so called PCR "hot-start"
procedures can be used, such as those employing heat
sensitive antibodies (Invitrogen, CA) to initially block the
activity of the polymerase.
25   Following the amplification step, optional steps
prior to length separation and detection improve the method's
signal to noise ratio. It is preferable to use the protocol
of Sec. 6.1.12.2.1 referred to as "Biotin bead clean-up."
This involves the use of a primer with a biotin (or capture
30 moiety) in the PCR amplification followed by binding to
streptavidin (or the capture moieties's binding partner) and
washing to remove certain reaction products. The single
strands denatured from the bound products are then further
analyzed. Further, single strands produced as a result of
35 linear amplification from singly cut fragments can be removed
by the use of single strand specific exonucleases. Mung Bean
exonuclease (Exo) or Exo I can be used, with Exo I preferred

because of its higher specificity for single strands. Mung bean is less preferred and even less preferred is S1 nuclease. Less preferably, the amplified products may be optionally concentrated by ethanol precipitation or column

5 separation.

Alternate PCR primers illustrated in Figure 10D can be advantageously used. In that figure, sample dsDNA 201 is illustrated after the RE/ligase reaction and after incubation at 72°C for 10 minutes but just prior to the PCR

10 amplification steps. dsDNA 201 has been cleaved by an RE recognizing subsequence 227 at position 221 producing overhang 202 and has been ligated to adapter primer strand 203. For definiteness and without limitation, a particular relative position between RE recognition subsequence 227 and

15 overhang 202 is illustrated. Other relative positions are known. The resulting DNA has been completed to a blunt ended double strand by completing strand 220 by incubation at 72°C for 10 minutes. Typically adapter primer strand 203 is used as the PCR primer.

20 Alternatively, strand 222, illustrated with its 5' end at the left, can be advantageously used. Strand 222 comprises subsequence 223, with the same sequence as strand 203; subsequence 224, with the same sequence as the RE overhang 202; subsequence 225, with a sequence consisting of

25 a remaining portion of RE recognition subsequence 227, if any; and subsequence 226 of P nucleotides. Length P is preferably from 1 to 6 and more preferably either 1 or 2. Subsequences 223 and 224 hybridize for PCR priming with corresponding subsequences of dsDNA 201. Subsequence 225

30 hybridizes with any remainder of recognition subsequence 227. Subsequence 226 hybridizes only with fragments 201 having complementary nucleotides in corresponding positions 228. When P is 1, primer 223 selects for PCR amplification 1 of the 4 possible dsDNAs 201 which may be present; and when P is

35 2, 1 of the 16 is selected. If 4 (or 16) primers 223 are synthesized, each with one of the possible (pairs of) nucleotides, and if the RE/ligase reactions mix is separated

in 4 (16) aliquots for use with one of these 4 (16) primers,
the 4 (16) PCR reactions will select for amplification only
one of the possible dsDNAs 201. Thus, these primers are
similar to phasing primers (European Patent Publication No.
5 O 534 858 A1, published Mar. 31, 1993).

The joint result of using primers 223 with
subsequence 226 in multiple PCR reactions after one RE/ligase
reaction is to extend the effective target subsequence from
the RE recognition subsequence by concatenating onto the
10 recognition subsequence a subsequence which is complementary
to subsequence 226. Thereby, many additional target
subsequences can be recognized while retaining the
specificity and exactness characteristic of the RE
embodiment. For example, REs recognizing 4 bp subsequences
15 can be used in such a combined reaction with an effective 5
or 6 bp target subsequence, which need not be palindromic.
REs recognizing 6 bp sequences can be used in a combined
reaction to recognize 7 or 8 bp sequences. Such effective
sequences are then used in the computer implemented design
20 and analysis methods subsequently described.

In a further enhancement, additional subsequence
information can be generated from adapters comprising primers
with specially placed Type IIS RE recognition subsequence
followed by digestion with the Type IIS RE and sequencing of
25 the generated overhang (in a SEQ-QEA™ embodiment). In a
preferred alternative, the Type IIS recognition subsequence
is placed so that the generated overhang is contiguous with
the original recognition subsequence of the RE that cut the
end to which the adapter hybridizes. In this embodiment, an
30 effective target subsequence is formed by concatenating the
sequence of the Type IIS overhang and the original
recognition sequence. In another alternative, the Type IIS
recognition sequence is placed so that the sequence of the
generated overhang is not contiguous with the original
35 recognition sequence. Here, the sequence of the overhang is
used as an third internal subsequence in the fragment. In
both cases, the additionally recognized subsequence is used

in the computer implemented experimental analysis methods to
increase the capability of determining the source sequence of
a fragment.   This enhancement is illustrated in Figures 23A-E
and is described in detail in Sec. 5.4.4 ("A SEQ-QEA™
5 Embodiment of a QEA™ Method).

A subsequent QEA™ method step is the separation by
length of the amplified, labeled, cut cDNA fragments and
observation of the length distribution. Lengths of the
sample of cut fragments will typically span a range from a
10 few tens of bp to perhaps 1000 bp. For this range standard
gel electrophoresis is capable of resolving separate
fragments which differ by three or more base pairs.
Knowledge of average fragment composition allows for
correction of composition induced small mobility differences
15 and permits resolution down to 1 bp. Any separation method
with adequate length resolution, preferably at least to three
base pairs in a 1000 base pair sequence, can also be used.
The length distribution is detected with means sensitive to
the primer labels. In the case of fluorochrome labels, since
20 multiple fluorochrome labels can be typically be resolved
from a single band in a gel, the products of one recognition
reaction with several REs or other recognition means or of
several separate recognition reaction can be analyzed in a
single lane. The detection apparatus resolution for
25 different labels limits the number of RE products that can be
simultaneously detected.

Preferred protocols for the specific RE embodiments
are described in detail in § 6.1.12.1 (entitled "The QEA™
Method Preferred RE Method").
30

5.4.3.1.   <u>FIRST ALTERNATIVE RE EMBODIMENT</u>
An alternative QEA™ method protocol performs
amplification prior to the RE/ligase step. After the
RE/ligase step, further amplification is performed.
35 Alternately, no further amplification is performed, and in
this case unwanted singly cut ends are removed as they are
not diluted by subsequent amplification.

- 154 -

Such removal is accomplished by first using primers that are labeled with a capture moiety. A capture moiety is a substance having a specific binding partner that can be affixed to a solid substrate. For example, suitable capture
5 moiety-binding partner pairs include, but are not limited to, biotin-streptavidin, biotin-avidin, a hapten (such as digoxigenin) and a corresponding antibody, or other removal means known in the art. For example, double stranded cDNA is PCR amplified using a set of biotin-labeled, arbitrary
10 primers with no net sequence preference. The result is partial cDNA sequences with biotin labels linked to both ends. The amplified cDNA is cut with REs and ligated to recognition moieties uniquely for each particular RE cut end. The RE/ligase step is performed by procedures identical to
15 those of the prior section in order to drive the RE digestion and recognition moiety ligation to completion and to prevent formation of concatamers and other unwanted ligation products. The recognition moieties can be the adapters previously described.

20   Next the unwanted singly cut fragments labeled with the capture moiety are removed by contacting them with the binding partner for the capture moiety affixed to a solid phase, followed by removal of the solid phase. For example, where biotin is the capture moiety, singly cut fragments can
25 be removed using streptavidin or avidin magnetic beads, leaving only doubly cut fragments that have RE-specific recognition moieties ligated to each end. These products are then analyzed, also as in the previous section, to determine the distribution of fragment lengths and RE cutting
30 combinations.

        Other direct removal means may alternatively be used in this embodiment of a QEA™ method. Such removal means include, but are not limited to, digestion by single strand specific nucleases or passage though a single strand specific
35 chromatographic column, for example, containing hydroxyapatite.

5.4.3.2.   SECOND ALTERNATIVE RE EMBODIMENT

A second alternative embodiment in conjunction with sufficiently sensitive detection means can eliminate altogether the amplification step. In the preferred RE

5 protocol, doubly cut fragments ligated to adapters are exponentially amplified, while unwanted, singly cut fragments are at best linearly amplified. Thus amplification dilutes the unwanted fragments relative to the fragments of interest. After ten cycles of amplification, for example, signals from

10 unwanted fragments are reduced to less than approximately 0.1% of the signals from the doubly cut fragments. Gene expression can then be quantitatively determined down to at least this level. A greater number of amplification cycles results in a greater relative dilution of signals from

15 unwanted singly cut fragments and, thereby, a greater sensitivity. But amplification bias and non-linearities interfere with the quantitative response of the method. For example, certain fragments will be preferentially PCR amplified depending on such factors as length and average

20 base composition.

For improved quantitative response, it is preferred to eliminate the bias accompanying the amplification steps. Then output signal intensity is linearly responsive to the number of input genes or sequences generating that signal.

25 In the case of common fluorescent detection means, a minimum of 6 x 10^{-18} moles of fluorochrome (approximately 10^5 molecules) is required for detection. Since one gram of cDNA contains about 10^{-6} moles of transcripts, it is possible to detect transcripts to at least a 1% relative level from

30 microgram quantities of mRNA. With greater mRNA quantities, proportionately rarer transcripts are detectable. Labeling and detection schemes of increased sensitivity permit use of less mRNA. Such a scheme of increased sensitivity is described in Ju et al., 1995, Fluorescent energy transfer

35 dye-labeled primers for DNA sequencing and analysis, *Proc. Natl. Acad. Sci. USA* 92:4347-4351. Single molecule detection means are about 10^5 times more sensitive than existing

fluorescent means (Eigen et al., 1994, *Proc. Natl. Acad. Sci. USA* 91:5740-5747).

To eliminate amplification steps, a preferred protocol uses a capture moiety separation means to directly

5 remove singly cut fragments from the desired doubly cut fragments.   Only the doubly cut fragments have a discrete length distribution dependent only on the input gene sequences.   The singly cut fragments have a broad non-diagnostic distribution depending on cDNA synthesis

10 conditions.   In this protocol, cDNA is synthesized using a primer labeled with a capture moiety, is circularized, cut with REs, and ligated to adapters.   Singly cut ends are then removed by contact with a solid phase to which a specific binding partner of the capture moiety is affixed.

15 Figs 12A, 12B, and 12C illustrate a second alternative RE protocol, which uses biotin as such a capture moiety for direct removal of the singly cut 3' and 5' cDNA ends from the RE/ligase mixture.   cDNA strands are amplified using, for example, a primer with a biotin molecule linked to

20 one of the internal nucleotides as one of the two primers in PCR.   Terminal phosphates are retained.

Figure 11A illustrates such a cDNA 401 with ends 407 and 408, poly(dA) sequence 402, poly(dT) primer 403 with biotin 404 attached.   405 is a recognition sequences for RE_1;

25 406 is a sequence for RE_2.   Fragment 409 is the cDNA sequence defined by these adjacent RE recognition sequences. Fragments 423 and 424 are singly cut fragments resulting from RE cleavages at sites 405 and 406.

Figure 12B illustrates that, next, the cDNA is

30 ligated into a circle. A ligation reaction using, for example, T4 DNA ligase is performed under sufficiently dilute conditions so that predominantly intramolecular ligations occur circularizing the cDNA, with a only a minimum of intermolecular, concatamer forming ligations.   Reaction

35 conditions favoring circularization versus concatamer formation are described in Maniatis, 1982, *Molecular Cloning A Laboratory Manual*, pp. 124-125, 286-288, Cold Spring

Harbor, NY.   Preferably, a DNA concentration of less than
approximately 1 μg/ml has been found adequate to favor
circularization.   Concatamers can be separated from
circularized single molecules by size separation using gel
5 electrophoresis, if necessary.   Figure 12B illustrates the
circularized cDNA.   Blunt end ligation occurred between ends
407 and 408.

       Then the circularized, biotin end labeled, cDNA is
cut with REs and ligated to adapters uniquely recognizing and
10 perhaps uniquely labeled for each particular RE cut.   The
RE/ligase step is performed by procedures as described in the
section hereinabove in order to drive RE digestion and primer
ligation to completion over formation of concatamers and
other unwanted ligation products.   Next, the unwanted singly
15 cut ends are removed using streptavidin or avidin magnetic
beads, leaving only doubly cut fragments that have RE-
specific recognition sequences ligated to each end.

       Figure 12C illustrates these latter steps.
Sequences 405 and 406 are cut by RE_1 and RE_2, respectively,
20 and adapters 421 and 422 specific for cuts by RE_1 and RE_2,
respectively are ligated onto the sticky ends.   Thereby,
fragment 409 is freed from the circularized cDNA and adapters
421 and 422 are ligated to it.   The remaining segment of the
circularized cDNA comprises singly cut ends 423 and 424 with
25 ligated adapters 421 and 422.   Both singly cut ends are
joined to the primer sequence 403 with attached biotin 404.
Removal is accomplished by contact with streptavidin or
avidin 420 which is fixed to substrate 425, perhaps
comprising magnetic beads.   The doubly cut labeled fragment
30 409 can now be simply separated from the singly cut ends
affixed to the substrate.   Thereby, separation of the singly
and doubly cut fragments is achieved.

       Signals from the uniquely labeled doubly cut ends
can be directly detected without any unwanted contamination
35 from signals from labeled singly cut ends.   Importantly,
since signals originate only from cDNA sequences originally
present in the sample, the detected signals will

- 158 -

quantitatively reflect cDNA sequence content and thus gene
expression levels. If the expression level is too low for
direct detection, the sample can be subjected to just the
minimum number of cycles of amplification, according to the
5 methods of Example 6.1.12.1 (entitled "Preferred QEA™ RE
Method"), to detect the gene or sequence of interest. For
example, the number of cycles can be as small as four to
eight without any concern of background contamination or
noise. Thus, in this embodiment, amplification is not needed
10 to suppress signals from singly cut ends, and preferred more
quantitative response signal intensities result.

5.4.4. A SEQ-QEA™ EMBODIMENT OF A QEA™ METHOD

SEQ-QEA™ is an alternative embodiment to the
15 preferred method of practicing a QEA™ method as described in
Sec. 5.4.3 ("RE Embodiments Of a QEA™ Method"). By the use
of recognition moieties, or adapters, comprising specially
constructed primers bearing a recognition site for a Type IIS
RE, a SEQ-QEA™ method is able to identify an additional 4-6
20 terminal nucleotides adjacent to the recognition site, or
recognition subsequence, of the RE initially cutting a
fragment. Thereby, the effective target subsequence is the
concatenation of the initial RE recognition subsequence and
the additional 4-6 terminal nucleotides, and has, therefore,
25 a length of at least from 8 to 12 nucleotides and preferably
has a length of at least 10 nucleotides. This longer
effective target subsequence is then used in the QEA™
analysis methods as described in Sec. 5.4.5 ("QEA™ Analysis
and Design Methods") which involve searching a database of
30 sequences to identify the sequence or gene from which the
fragment derived. The longer effective target subsequence
increases the capability of these methods to determine a
unique source sequence for a fragment.

In this section, for ease of description and not by
35 way of limitation, first shall be described Type IIS REs,
next the specially constructed primers, and then the

additional method steps of a SEQ-QEA™ method used to
recognize the additional nucleotides.

A Type IIS RE is a restriction endonuclease enzyme
which cuts a dsDNA molecule at locations outside of the
5 recognition site of the Type IIS RE (Szybalski et al., 1991,
Gene 100:13-26). Figure 23C illustrates Type IIS RE 2331
cutting dsDNA 2330 outside of its recognition site, which is
recognition subsequence 2320, at locations 2308 and 2309.
The Type IIS RE preferably generates an overhang by cutting
10 the two dsDNA strands at locations differently displaced away
on the two strands from the recognition subsequence.
Although the recognition subsequence and the displacement(s)
to the cutting site(s) are determined by the RE and are
known, the sequence of the generated overhang is determined
15 by the dsDNA cut, in particular by its nucleotide sequence
outside of the Type IIS recognition region, and is, at first,
unknown. Thus, in a SEQ-QEA™ embodiment, the overhangs
generated by the Type IIS REs are sequenced. Table 9 in Sec.
6.1.12.5 ("Preferred Reactants for SEQ-QEA™ Methods") lists
20 several Type IIS REs adaptable for use in a SEQ-QEA™ method
and their relevant characteristics, including their
recognition subsequences on both DNA strands and the
displacements from these recognition subsequences to the
respective cutting sites. It is preferable to use REs of
25 high specificity and generating an overhang of at least 4 bp
displaced at least 4 or 5 bp beyond the recognition
subsequence in order to span the remaining recognition
subsequence of the RE that initially cut the fragment. FokI
and BbvI are most preferred Type IIS REs for a SEQ-QEA™
30 method.

Next, the special primers, and the special linkers
if needed, which hybridize to form the adapters for SEQ-QEA™,
have, in additional to the structure previously described in
Sec. 5.4.3 ("RE Embodiments Of a QEA™ Method"), a Type IIS
35 recognition subsequence whose placement is important in order
that the overhang generated by the Type IIS enzyme be
contiguous to the initial target end subsequence. The

placement of this additional subsequence is described with
reference to Figures 23A-E, which illustrate steps in a SEQ-
QEA™ embodiment. Fig 23B schematically illustrates dsDNA
2302, which is a fragment cut from an original sample

5 sequence on one end by a first initial RE and on the other
end by a different second initial RE, with adapters fully
hybridized but prior to primer ligation. Thus, linker strand
2311 has hybridized to primer strand 2312 and to the 5'
overhang generated by the first initial RE, and now fixes

10 primer 2312 adjacent to fragment 2302 for subsequent
ligation. Primer 2312 has recognition subsequence 2320 for
Type IIS RE 2321. Linker 2311, to the extent it overlaps and
hybridizes with recognition subsequence 2320, has
complementary recognition subsequence 2321. Additionally,

15 primer 2312 preferably has a conjugated label moiety 2334,
e.g. a fluorescent FAM moiety. Similarly, linker strand 2313
has hybridized to primer strand 2314 and to the 5' overhang
generated by the second initial RE. Primer 2314 preferably
has a conjugated capture moiety 2332, e.g. a biotin moiety,

20 and a release means represented by subsequence 2323 (to be
described subsequently). Primer 2312 is also called the "cut
primer," and primer 2314 the "capture primer."

          Subsequence 2304 terminating at nucleotide 2307 in
Figure 23B is the portion of the recognition subsequence of

25 the first initial RE remaining after its cutting of the
original sample sequence. The placement of the Type IIS RE
recognition subsequence is determined by the length of this
subsequence. Figure 23A schematically illustrates how the
length of subsequence 2304 is determined by properties of the

30 first initial RE. The first RE is chosen to be of a type
that recognizes subsequence 2303, terminating with nucleotide
2307, of sample dsDNA 2301, and that cuts the two strands of
dsDNA 2301 at locations 2305 that are located within
recognition subsequence 2303. In order that the first RE

35 recognize a known target subsequence, it is highly preferable
that subsequence 2303 be entirely determined by the first RE
and be without indeterminate nucleotides. As a result of

- 161 -

this cutting, overhang subsequence 2306 is generated and has
a known sequence, since it is entirely within the determined
recognition subsequence 2303. Thereby, subsequence 2304, the
portion of the recognition subsequence 2303 remaining on a
5 fragment cut by the first RE, has a length not less than the
length of overhang 2306 and is typically longer. Typically
and preferably, subsequence 2303 is of length 6 and is
palindromic; locations 2305 are symmetrically placed in
subsequence 2303; and overhang 2306 is of length 4.
10 Therefore, the typical length of the remaining portion 2304
of the recognition subsequence 2303 is of length 5. In cases
where shorter recognition subsequences 2303 are preferably,
the remaining portion 2304 will have a corresponding length.
          The preferred placement of Type IIS recognition
15 sequence 2320 is now described with reference to Figure 23C,
which schematically illustrates dsDNA 2330, which derives
from dsDNA 2302 of Figure 23B after the further steps of
primer ligation, PCR amplification with primers 2312 and
2314, binding of capture moiety 2332 to binding partner 2333
20 affixed to a solid-phase substrate, and then binding of Type
IIS RE 2331 to its recognition subsequence 2320. Subsequence
2322 is the subsequence between recognition subsequence 2320
and the end of primer 2312 at location 2305. Type IIS RE is
illustrated cutting dsDNA 2330 at nucleotide locations 2308
25 and 2309 and, thereby, generating an exemplary 5' overhang
2324 between these locations. For this overhang to be
contiguous with the remaining portion 2304 of initial target
end subsequence 2303, nucleotide 2309 is adjacent to
nucleotide 2307 terminating subsequence 2304. Therefore,
30 Type IIS recognition sequence 2320 is preferably placed on
primer 2312 such that the length of subsequence 2304 plus the
length of subsequence 2322 equals the distance of closest
cutting of Type IIS RE 2331. For example, in the case of
FokI, since the closest cutting distance is 9 and the typical
35 length of subsequence 2304 is 5, its recognition sequence is
preferably placed 4 bp from the end of primer 2312. In the
case of BbvI, since the closest cutting distance is 8, its

recognition sequence is preferably placed 3 bp from the end
of primer 2312.

Finally, Figure 23D schematically illustrates dsDNA
2330 after cutting by Type IIS RE 2331. dsDNA has 5'
5 overhang 2324 between and including nucleotides 2308 and
2309, where the Type IIS RE cut dsDNA 2330 of Figure 23C.
This overhang is contiguous with former subsequence 2304, the
remaining portion of the recognition subsequence of the first
RE, which has been cut off. The shorter strand has primer
10 2314 including release means represented by subsequence 2323.
dsDNA 2330 remains bound to the solid-phase support through
capture moiety 2332 and binding partner 2324. The absence of
label moiety 2334 can be used to monitor the completeness of
cutting by Type IIS RE 2331. The label moiety also
15 advantageously assists in the determination of the length of
dsDNA 2330.

The QEA™ method is also adaptable to other less
preferable placements of recognition sequence 2320. If
recognition sequence 2320 is placed closer to the 3' end of
20 primer 2312 than the optimal and preferable distance, the
overhang produced by Type IIS RE 2331 is not contiguous with
recognition subsequence 2303 of the first RE, and a
contiguous effective target subsequence is not generated. In
this case, optionally, the determined sequence of the Type
25 IIS RE generated overhang can be used as third internal
subsequence information in QEA™ experimental analysis methods
in order to further resolve the source sequence of fragment
2302, if necessary. If recognition sequence 2320 is placed
further from the 3' end of the cut primer than the optimal
30 and preferable distance, the overhang produced by Type IIS RE
overlaps with recognition subsequence 2303 of the first RE.
In this case, the length of the now contiguous effective
target subsequence is less than the sum of the lengths of the
Type IIS overhang and the first RE recognition subsequence.
35 Effective target end subsequence information is, thereby,
lost. In case recognition sequence 2310 is placed further

- 163 -

from the 3' end than the distance of furthest cutting, no
additional information is obtained.

Primer 2314 also has certain additional structure.
First additional structure is capture moiety 2332 conjugated
5 near or to the 5' end of primer 2314. The capture moiety
cooperates with a corresponding binding partner affixed to a
solid support, an attachment means, to immobilize dsDNA 2330.
Biotin/streptavidin are the preferred capture moiety/binding
partner pair, which are used in the following description
10 without limitation to this invention. This embodiment is
adaptable to any cooperating pair of capture moiety and
binding partner that remain bound under DNA denaturing
conditions. Several such pairs have been previously
described.

15   A second additional structure is a release means
represented as subsequence 2323 of primer 2314. The release
means allows controlled release of strand 2335 of Figure 23D
from the capture moiety/binding partner complex. This
alternative is adaptable to any such controlled release
20 means. Two such means are preferable. First, subsequence
2323 can be one or more uracil nucleotides. In this case,
digestion with uracil DNA glycosylase (UDG) and subsequent
hydrolysis of the sugar backbone at an alkaline pH releases
strand 2335. Second, subsequence 2323 can be the recognition
25 subsequence of an RE which cuts extremely rarely if at all in
the sequences of the sample. A preferred RE of this sort is
AscI, which has an 8 bp recognition sequence that rarely, if
ever, occurs in mammalian DNA, and is active at the ends of
molecules. In this case, digestion with the RE, $i.e.$ AscI,
30 releases strand 2335. These release means are particularly
useful in the case of biotin-streptavidin, which form a
complex that is difficult to dissociate.

Table 10 of Sec. 6.1.12.5 ("Preferred Reactants for
SEQ-QEA™ Methods") lists exemplary primers, linkers, and
35 associated REs, for the preferred implementation of SEQ-QEA™
in which contiguous effective target end subsequences are
formed. This description has illustrated the generation of a

- 164 -

5' Type IIS generated overhang. Primers can equally be constructed to generate a less preferable 3' overhang by using a Type IIS whose closest cutting distance is on the 3' strand, rather than on the 5' strand.

5   Finally, the method steps of SEQ-QEA™ are now described. SEQ-QEA™ comprises, first, practicing the RE/ligase embodiment of QEA™ using the special primers and linkers previously described followed, second, by certain additional steps specific to SEQ-QEA™. More detailed

10 exemplary reaction protocols are found in the accompanying examples in Sec. 6 ("Examples"). The protocols of Sec. 6.1.12.1 ("Preferred QEA™ RE Method") are preferred for performing a QEA™ method, and the protocols of Sec. 6.1.12.2 ("Preferred Methods Of A SEQ-QEA™ Embodiment") are preferred

15 for performing the additional steps specific to SEQ-QEA™. Figures 23B-E illustrate various steps in a SEQ-QEA™ method. Figure 23B illustrates a fragment from a sample sequence digested by two different REs and just prior to primer ligation. Figure 23C illustrates a sample sequence after

20 primer ligation, chain blunt-ending, and PCR amplification. These QEA™ steps are preferably performed according to the alternative described in Sec. 5.4.3 ("RE Embodiments Of a QEA™ Method"), but can alternatively be performed by any RE/ligase alternative. The additional steps unique to SEQ-

25 QEA™ include, first, binding the amplified fragments to a solid-phase support, also illustrated in Figure 23C, second, washing the bound fragments, and third, digesting the bound fragments by the Type IIS RE corresponding to primer 2312 used. The Type IIS digestion is preferably performed with

30 reaction conditions suitable to achieve complete digestion, which can be checked by insuring the absence of optional label moiety 2334 after washing the bound, digested sequences. Figure 23D illustrates dsDNA fragments 2330 remaining after complete digestion by the Type IIS RE.

35 Before Type IIS digestion, an aliquot of the bound, amplified RE/ligase reaction products is denatured and the supernatant, containing the labeled 5' strands, are separated according to

length by, e.g., gel electrophoresis, in order to determine
the length of each fragment doubly cut by different REs as in
the previous QEA™ embodiments.

The subsequent additional SEQ-QEA™ step is
5 sequencing of overhang 2324. This can be done in any manner
known in the art. In a preferred embodiment suitable for
lower fragment quantities, an alternative, herein called
phasing QEA™, can be used to sequence this overhang. A
phasing QEA™ method depends on the precise sequence
10 specificity with which RE/ligase reactions recognize short
overhangs, in this case the Type IIS generated overhang.
Figure 23E illustrates a first step of this alternative in
which a QEA™ method adapter, which is comprised of primer
2351 with label moiety 2353 and linker 2350, has hybridized
15 to overhang 2324 in Type IIS digested fragment 2330 bound to
a solid-phase support. By way of example only, overhang 2324
is here illustrated as being 4 bp long. In this alternative,
special phasing linkers are used. For each nucleotide
position of overhang 2324, e.g. position 2354, 4 pools of
20 linkers 2350 are prepared. All linkers in each pool have one
fixed nucleotide, i.e. one of either A, T, C, or G, at that
position, e.g. position 2355, while random nucleotides in all
combinations are present at the other three positions. For
each nucleotide position of the overhang, four RE/ligase
25 reactions are performed according to QEA™ protocols, one
reaction using linkers from one of the four corresponding
pools. Linkers from only one pool, that having a nucleotide
complementary to overhang 2324 at position 2354, hybridize
without error, and only these linkers can cause ligation of
30 primer 2351 to the 5′ strand of fragment 2330. When the
results of the four RE/ligase reactions are denatured and
separated according to length, only one reaction of the four
can produce labeled products at a length corresponding to the
length of fragment 2330, namely the reaction with linkers
35 complementary to position 2354 of overhang 2324. Thereby, by
performing four RE/ligase reactions for each nucleotide
position of overhang 2324, this overhang can be sequenced.

Optionally, the products of these four RE/ligase reactions
can be further PCR amplified. In a further option, if
linkers 2350 comprise subsequence 2356 that is uniquely
related to the fixed nucleotide in subsequence 2352 and if
5 four separately and distinguishably labeled primers 2351
complementary to these unique subsequences are used, all four
RE/ligase for one overhang position reactions can be
simultaneously performed in one reaction tube. With this
overhang sequencing alternative, release means 2323 can be
10 omitted from primer 2314.

In an alternate embodiment, sequencing of a 5'
overhang can be done by standard Sanger reactions. Thus
strand 2335 is elongated by a DNA polymerase in the presence
of labeled ddNTPs at a relatively high concentration to dNTPs
15 in order to achieve frequent incorporation in the short 4-6
bp elongation. Partially elongated strands 2335 are released
by denaturing fragment 2330, washing, and then by causing
release means 2323 to release strands 2335 from the capture
moiety bound to the solid phase support. The released,
20 partially elongated strands are then separated by length,
e.g., by gel electrophoresis, and the chain terminating ddNTP
is observed at the length previously observed for that
fragment. In this manner, the 4-6 bp overhang 2324 of each
fragment can be quickly sequenced.

25 The effective target subsequence information,
formed by concatenating the sequence of the Type IIS overhang
to the sequence of the recognition subsequence of the first
RE, is then input into QEA™ Experimental Analysis methods,
and is used as a longer target subsequence in order to
30 determined the source of the fragment in question. This
longer effective target subsequence information preferably
permits exact and unique sample sequence identification.

5.4.5. QEA™ ANALYSIS AND DESIGN METHODS

35 Described hereinbelow are two groups of computer
methods: first, methods for the QEA™ method experimental
design; and second, methods for the QEA™ method experimental

analysis. Although, logically, design precedes analysis, the
methods of experimental design depend on basic methods
described herein as part of experimental analysis.
Consequently, experimental analysis methods are described
5 first.

In the following, descriptions are often cast in
terms of the preferred QEA™ method embodiment, in which REs
are used to recognize target subsequences. However, such
description is not limiting, as all the methods to be
10 described are equally adaptable to all QEA™ method
embodiments.

Further, the following descriptions are directed to
the currently preferred embodiments of these methods.
However, it will be readily apparent to those skilled in the
15 computer and simulation arts that many other embodiments of
these methods are substantially equivalent to those described
and can be used to achieve substantially the same results.
The QEA™ methods comprise such alternative implementations as
well as its currently preferred implementation.
20

5.4.5.1. QEA™ EXPERIMENTAL ANALYSIS METHODS

The analysis methods comprise, first, selecting a
database of DNA sequences representative of the DNA sample to
be analyzed, second, using this database and a description of
25 the experiment to derive the pattern of simulated signals,
contained in a database of simulated signals, which will be
produced by DNA fragments generated in the experiment, and
third, for any particular detected signal, using the pattern
or database of simulated signals to predict the sequences in
30 the original sample likely to cause this signal. Further
analysis methods present an easy to use user interface and
permit determination of the sequences actually causing a
signal in cases where the signal may arise from multiple
sequences, and perform statistical correlations to quickly
35 determine signals of interest in multiple samples.

The first analysis method is selecting a database
of DNA sequences representative of the sample to be analyzed.

In one use of a QEA™ method, the DNA sequences to be analyzed
will be derived from a tissue sample, typically a human
sample examined for diagnostic or research purposes. In this
use, database selection begins with one or more publicly
5 available databases which comprehensively record all observed
DNA sequences. Such databases are GenBank from the National
Center for Biotechnology Information (Bethesda, MD), the EMBL
Data Library at the European Bioinformatics Institute
(Hinxton Hall, UK) and databases from the National Center for
10 Genome Research (Santa Fe, NM). However, as any sample of a
plurality of DNA sequences of any provenance can be analyzed
by QEA™ methods, any database containing entries for the
sequences likely to be present in such a sample to be
analyzed is usable in the further steps of the computer
15 methods.

Figure 13A illustrates the preferred database
selection method starting from a comprehensive tissue derived
database. Database 1001 is the comprehensive input database,
having the exemplary flat-file or relational structure 1010
20 shown in Figure 13B, with one row, or record, 1014 for each
entered DNA sequence. Column, or field, 1011 is the
accession number field, which uniquely identifies each
sequence in database 1001. Most such databases contain
redundant entries, that is multiple sequence records are
25 present that are derived from one biological sequence.
Column 1013 is the actual nucleotide sequence of the entry.
The plurality of columns, or fields, represented by 1012
contain other data identifying this entry including, for
example whether this is a cDNA or gDNA sequence, if cDNA,
30 whether this is a full length coding sequence or a fragment,
the species origin of the sequence or its product, the name
of the gene containing the sequence, if known, etc. Although
shown as one file, DNA sequence databases often exits in
divisions and selection from all relevant divisions is
35 contemplated by a QEA™ method. For example, GenBank has 15
different divisions, of which the EST division and the
separate database, dbEST, that contain expressed sequence

- 169 -

tags ("EST") are of particular interest, since they contain expressed sequences.

From the comprehensive database, all records are selected which meet criteria for representing particular
5 experiments on particular tissue types. This is accomplished by conventional techniques of sequentially scanning all records in the comprehensive database, selecting those that match the criteria, and storing the selected records in a selected database.

10 The following are exemplary selection methods. To analyze a genomic DNA sample, database 1001 is scanned against criteria 1002 for human gDNA to create selected database 1003. To analyze expressed genes (cDNA sequences), several selection alternatives are available. First, a
15 genomic sequence can be scanned in order to predict which subsequences (exons) will be expressed. Thus selected database 1005 is created by making selections according to expression predictions 1004. Second, observed expressed sequences, such as cDNA sequences, coding domain sequences
20 ("CDS"), and ESTs, can be selected 1006 to create selected database 1007 of expressed sequences. Additionally, predicted and observed expressed sequences can be combined into another, perhaps more comprehensive, selected database of expressed sequences. Third, expressed sequences
25 determined by either of the prior methods may be further selected by any available indication of interest 1008 in the database records to create more targeted selected database 1009. Without limitation, selected databases can be composed of sequences that can be selected according to any available
30 relevant field, indication, or combination present in sequence databases.

The second analysis method uses the previously selected database of sequences likely to be present in a sample and a description of an intended experiment to derive
35 a pattern of the signals which will be produced by DNA fragments generated in the experiment. This pattern can be stored in a computer implementation in any convenient manner.

In the following, without limitation, it is described as
being stored as a table of information. This table may be
stored as individual records or by using a database system,
such as any conventionally available relational database.

5 Alternatively, the pattern may simply be stored as the image
of the in-memory structures which represent the pattern.

A QEA™ method experiment comprises several
independent recognition reactions applied to the DNA sample
sequences, where in each of the reactions labeled DNA

10 fragments are produced from sample sequences, the fragments
lying between certain target subsequences in a sample
sequence. The target subsequences can be recognized and the
fragments generated by the preferred RE embodiments of the
QEA™ method. The following description is focused on the RE

15 embodiments.

Figure 14 illustrates an exemplary description 1100
of a preferred QEA™ method embodiment. Field 1101 contains a
description of the tissue sample which is the source of the
DNA sample. For example, one experiment could analyze a

20 normal prostrate sample; a second otherwise identical
experiment could analyze a prostrate sample with premalignant
changes; and a third experiment could analyze a cancerous
prostate sample. Differences in gene expression between
these samples, particularly among interacting proteins

25 detected according to the method of the invention, then
relate to the progress of the cancer disease state. Such
samples could be drawn from any other human cancer or
malignancy.

Major rows 1102, 1105, and 1109 describe the

30 separate individual recognition reactions to which the DNA
from tissue sample 1101 is subjected. Any number of
reactions may be assembled into an experiment, from as few as
one to as many as there are pairs of available recognition
means to recognize subsequences. Figure 14 illustrates 15

35 reactions. For example, reaction 1 specified by major row
1102 generates fragments between target subsequences which
are the recognition sites of restriction endonucleases 1 and

2 described in minor rows 1103 and 1104. Further, the RE1 cut end is recognized by a labeling moiety labeled with LABEL1, and the RE2 end is recognized by LABEL2. Similarly, reaction 15, 1109, utilizes restriction endonucleases 36 and
5 37 labeled with labels 3 and 4, minor rows 1110 and 1111, respectively.

Major row 1105 describes a variant QEA™ method reaction using three REs and a separate probe. As described, many REs can be used in a single recognition reaction as long
10 as a useful fragment distribution results. Too many REs results in a compressed length distribution. Further, probes for target subsequences that are not intended to be labeled fragment ends, but rather occur within a fragment, can be used. For a further example, a labeled probe added after the
15 QEA™ method PCR amplification step (if present in a given embodiment), a post PCR probe, can recognize subsequences internal to a fragment and thereby provide an additional signal which can be used to discriminate between two sample sequences which produce fragments of the same length and end
20 sequence which otherwise have differing internal sequences. For another example, a probe added before a QEA™ method PCR step and which cannot be extended by DNA polymerase will prevent PCR amplification of those fragment containing the probe's target subsequences. If PCR amplification is
25 necessary to generate detectable signals (in a given embodiment), such a probe will prevent the detection of such a fragment. The absence of a fragment may make a previously ambiguous detected band now unambiguous. Such PCR disruption probes can be PNA oligomers or degenerate sets of DNA
30 oligomers, modified to prevent polymerase extension (e.g., by incorporation of a dideoxynucleotide at the 3' end).

In certain QEA™ method embodiments an effective target subsequence is available that is longer than the recognition subsequence of the cutting RE. In these cases,
35 the effective target subsequence is to be used in the analysis and design methods in place of the cutting RE recognition subsequence in order to obtain extra specificity.

One such embodiment is a SEQ-QEA™ method, wherein an overhang generated by a Type IIS RE is sequenced to obtain a longer target end subsequence. Another such embodiment involves the use of alternative phasing PCR primers. In this case, their
5 extra recognition subsequences and labeling are described in rows dependent to the RE/ligase reaction whose products they are used to amplify.

Next, Figure 15A illustrates, in general, that from the database selected to best represent the likely DNA
10 sequences in the sample analyzed, 1201, and the description of the QEA™ method experiment, 1202, the simulation methods, 1203, determine a pattern of simulated signals stored in a simulated database, 1204, that represents the results of the QEA™ method experiment. The experimental simulation
15 generates the same fragment lengths and end subsequences from the input database that will be generated in an actual experiment performed on the same sample of DNA sequences.

Alternately, the simulated pattern or database may not be needed, in which case the DNA database is searched
20 sequence by sequence, mock digestions are performed and compared against the input signals. A simulated database is preferable if several signals need to be searched or if the same QEA™ method experiment is run several times. Conversely, the simulated database can be dispensed with when
25 few signals from a few experiments need to searched. A quantitative statement of when the simulated database is more efficient depends upon an analysis of the costs of the various operations and the size of DNA database, and can be performed as is well known in the computer arts. Without
30 limitation, in the following the simulated database is described

Figure 15B illustrates an exemplary structure for the simulated database. Here, the simulated results of all the individual recognition reactions defined for the
35 experiment are gathered into rectangular table 1210. The QEA™ method is equally adaptable to other database structures containing equivalent information; such an equivalent

structure would be one, for example, where each reaction was placed in a separate table. The rows of table 1210 are indexed by the lengths of possible fragments. For example, row 1211 contains fragments of length 52. The columns of

5 table 1210 are indexed by the possible end subsequences and probe hits, if any, in a particular experimental reaction. For example, columns 1212, 1213, and 1214 contain all fragments generated in reaction 1, R1, which have both end subsequences recognized by RE1, one end subsequence

10 recognized by RE1 and the other by RE2, and both end subsequences recognized by RE2, respectively. Other columns relate to other reactions of the experiment. Finally, the entries in table 1210 contain lists of the accession numbers of sequences in the database that give rise to a fragment

15 with particular length and end subsequences. For example, entry 1215 indicates that only accession number.A01 generates a fragment of length 52 with both end subsequences recognized by RE1 in R1. Similarly, entry 1216 indicates that accession numbers A01 and S003 generate a fragment of length 151 with

20 both end subsequences recognized by RE3 in reaction 2.

In alternative embodiments, the contents of the table can be supplemented with various information. In one aspect, this information can aid in the interpretation of results produced by the separation and detection means used.

25 For example, if separation is by electrophoresis, then the detected electrophoretic DNA length can be corrected to obtain the true physical DNA length. Such corrections are well known in the electrophoretic arts and depend on such factors as average base composition and fluorochrome labels.

30 One commercially available package for making these corrections is Gene Scan Software from Applied Biosystems, Inc. (Foster City, CA). In this case, each table entry for a fragment can contain additionally average base composition, perhaps expressed as percent G+C content, and the

35 experimental definition can include primer average base composition and fluorochrome label used. For a further example, if separation is by mass spectroscopy or similar

method, the additional information can be the molecular
weight of each fragment and perhaps a typically fragmentation
pattern. Use of other separation and detection means can
suggest the use of other appropriate supplemental data.

5   Where phasing primers are used, supplemental
columns are used with RE pairs in order to further identify
the effective target subsequence. A similar method can be
employed to take account of the SEQ-QEA™ method additional
subsequence information. In this latter case, the additional
10 information is not available until after the QEA™ method
experiment is performed.

         Before describing how this simulated database is
generated, it is useful first to describe how this database
is used to predict experimental results. Returning to Figure
15 14, labels are used to detect binding reaction events by
subsequence recognition means to the target DNA, to allow
detection after separation of the fragments by length. In an
embodiment using fluorescent detection means, these labels
are fluorochromes covalently attached to the primer strands
20 of the adapters, as previously described, or to hybridization
probes, if any. Typically, all the fluorochrome labels used
in one reaction are simultaneously distinguishable so that
fragments with all possible combinations of target
subsequences can be fluorescently distinguished. For
25 example, fragments at entry 1217 in table 1210 (Figure 15B)
occur at length 175 and present simultaneous fluorescent
signals LABEL1 and LABEL2 upon stimulation, since these are
the labels used with adapters which recognize ends cuts by
RE1 and RE2 respectively. For a further example, in reaction
30 2, major row 1105 of experimental definition 1100 (Figure
14), a fragment with ends cut by RE2 and RE3 and hybridizing
with probe P will present simultaneous signals LABEL2,
LABEL3, and LABEL4. Where effective target subsequences are
constructed with the SEQ-QEA™ method or alternative phasing
35 primers, this lookup is appropriately modified.

         Other labelings are within the scope of the QEA™
method. For example, a certain group of target subsequences

can be identically labeled or not labeled at all, in which
case the corresponding group of fragments are not
distinguishable. In this case, if RE1 and RE3 end
subsequences were identically labeled in table 1210 (Figure
5 15B), a fragment of length 151 may be generated by sequence
T163, A01, or S003, or any combination of these sequences.
In the extreme, if silver (Ag) staining of an electrophoresis
gel is used in an embodiment to detect separated fragments,
then all bands will be identically labeled and only band
10 lengths can be distinguished within one electrophoresis lane.
        Thus the simulated database together with the
experimental definition can be used to predict experimental
results. If a signal is detected in a recognition reaction,
say Rn, whose end labelings are LABEL1 and LABEL2 and whose
15 representation of length is corrected to physical length in
base pairs of L, the length L row of the simulated database
is retrieved and it is scanned for Rn entries with the
detected subsequence labeling, by using the column headings
indicating observed subsequences and the experimental
20 definition indicating how each subsequence is labeled. If no
match is found, this fragment represents a new gene or
sequence not present in the selected database. If a match is
found, then this fragment, in addition to possibly being a
new gene or sequence, can also have been generated by those
25 candidate sequences present in the table entry(ies) found.
        The simulated database lookup is described herein
as using the physical length of a detected fragment. In
cases where the separation and detection means returns an
approximation to the true physical fragment length, lookup is
30 augmented to account for such as approximation. For example,
electrophoresis, when used as the separation means, returns
the electrophoretic length, which depending on average base
composition and labeling moiety is typically within 10% of
the physical length. In this case database lookup can search
35 all relevant entries whose physical length is within 10% of
the reported electrophoretic length, perform corrections to
obtain electrophoretic length, and then check for a match

- 176 -

with the detected signal. Alternative lookup implementations
are apparent, one being to precompute the electrophoretic
length for all predicted fragments, construct an alternate
table index over the electrophoretic length, and then
5 directly lookup the electrophoretic length. Other separation
and detection means can require corresponding augmentations
to lookup to correct for their particular experimental biases
and inaccuracies. It is understood that where database
lookup is referred to subsequently, either simple physical
10 lookup or augmented lookup is meant as appropriate.

If matched candidate database sequences are found,
then the selected database can be consulted to determine
other information concerning these sequences, for example,
gene name, tissue origin, chromosomal location, etc. If an
15 unpredicted fragment is found, this fragment can be
optionally retrieved from the length separation means, cloned
or sequenced, and used to search for homologues in a DNA
sequence database or to isolate or characterize the
previously unknown gene or sequence. In this manner, the
20 QEA™ method can be used to rapidly discover and identify new
genes.

The QEA™ computer methods are also adaptable to
other formats of an experimental definition. For example,
the labeling of the target subsequence recognition moieties
25 can be stored in a table separate from the table defining the
experimental reactions.

Now turning to the methods by which the simulated
database is generated, Figure 16 illustrates a basic method,
termed herein mock fragmentation, which takes one sequence
30 and the definition of one reaction of an experiment and
produces the predicted results of the reaction on that
sequence. Generation of the entire simulated database
requires repetitive execution of this basic method.

Turning first to a description of mock
35 fragmentation, the method commences at 1301 and at 1302 it
inputs the sequence to be fragmented and the definition of
the fragmentation reaction, in the following terms: the

target end subsequences RE1 ... REn, where n is typically 2
or 3, and the subsequences to be recognized by third
subsequence probes, P1 ... Pn, where n is typically 0 or 1.
Note that post PCR disruption probes act as unlabeled end
5 subsequences and are so treated for input to this method.
The operation of the method is illustrated by example in
Figure 17A-F for the case RE1, RE2 and P1.

     At step 1303, for each target end subsequence, the
method makes a "vector of ends", which has elements which are
10 pairs of nucleotide positions along the sequence, each pair
being labeled by the corresponding end subsequence. For
embodiments where end subsequences are recognized by
hybridizing oligonucleotides, the first member of each pair
is the beginning of a target end subsequence and the second
15 member is the end of a target end subsequence. For
embodiments where target end subsequences are recognized by
restriction endonucleases, the first member of each pair is
the beginning of the overhang region that corresponds to the
RE recognition subsequence and the second member is the end
20 of that overhang region. It is preferred to use REs that
generate 4 bp overhangs. The actual target end subsequences
are the RE recognition sequences, which are preferably 4-8 bp
long.

     This vector is generated by a string operation
25 which compares the target end subsequence in a 5' to 3'
direction against the input sequence and seeks string
matches, that is the nucleotides match exactly. Where
effective target subsequences are formed by using the SEQ-
QEA™ method or alternative phasing primers, it is the
30 effective subsequences that are compared. This can be done
by simply comparing the end subsequence against the input
sequence starting at one end and proceeding along the
sequence one base at time. However, it is preferable to use
a more efficient string matching algorithm, such as the
35 Knuth-Morris-Pratt or the Boyer-Moore algorithms. These are
described with sample code in Sedgewick, 1990, *Algorithms in
C*, chap. 19, Addison-Wesley, Reading, MA.

In embodiments of the QEA™ method wherein target subsequence are recognized with accuracy, such as the RE embodiments, the comparison of target subsequence against input sequence should be exact, that is the bases should
5 match in a one-to-one manner. In embodiments where target subsequences are less accurately recognized, the string match should be done in a less exact, or fuzzy, manner. In this case the string operation, which generates the vector of ends, should accept partial T-n matches as well as exact
10 matches. In this, the string operations generate the false positive matches expected from the experiments and permit these fragments to be identified. Ambiguity in the simulated database, however, increases, since more fragments leads to a greater chance of fragments of identical length and end
15 labels.

Figure 17A illustrates end vectors 1401 and 1402, comprising three and two ends, respectively, generated by RE1 and RE2, which are for this example assumed to be REs with a 4 bp overhang. The first overhang in vector 1401 occurs
20 between nucleotide 10 and 14 in the input sequence.

Step 1304 of Figure 16 merges all the end vectors for all the end subsequences and sorts the elements on the position of the end. Vector 1404 of Figure 17B illustrates the result of this step for example end vectors 1401 and
25 1402.

Step 1305 of Figure 16 then creates the fragments generated by the reaction by selecting the parts of the full input sequence that are delimited by adjacent ends in the merged and sorted end vector. Since the experimental
30 conditions in conducting the QEA™ method should be selected such that target end subsequence recognition is allowed to go to completion, all possible ends are recognized. For the restriction endonuclease embodiments, the cutting and ligase reactions should be conducted such that all possible RE cuts
35 are made and to each cut end a labeled primer is ligated. These conditions insure that no fragments contain internal unrecognized target end subsequences and that only adjacent

ends in the merged and sorted vector define generated fragments.

Where additional information is needed for simulated database entries to adapt to inaccuracies in
5 particular separation and detection means, such information can be collected at this step. For example, in the case of electrophoretic separation, fragment sequence can be determined and percent G+C content computed and entered in the database along with the fragment accession number.
10   For the PCR embodiments, the fragment length is the difference between the end position of the second end subsequence and the start position of the first end subsequence. For RE embodiments, the fragment length is the difference between the start position of the second end
15 subsequence and the start position of the first end subsequence plus twice the primer length (48 in the preferred primer embodiment).

Figure 17C illustrates the exemplary fragments generated, each fragment being represented by a 4 member
20 tuple comprising: the two end subsequences, the length, and an indicator whether the third subsequence probe binds to this fragment. In Figure 17C the position of this indicator is indicated by a '*'. Fragment 1408 is defined by ends 1405 and 1406, and fragment 1409 by ends 1406 and 1407. There is
25 no fragment defined by ends 1405 and 1407 because the intermediate end subsequence is recognized and either fully cut in an RE embodiment or used as a fragment end priming position in a PCR embodiment. For simplicity, the fragment lengths are illustrated for the RE embodiment without the
30 primer length addition.

Step 1306 of Figure 16 checks if a hybridization probe is involved in the experiment. If not, the method skips to step 1309. If so, step 1307 determines the sequence of the fragment defined in step 1305. Figure 17D illustrates
35 that the fragment sequences for this example are the nucleotide sequences within the input sequence that are between the indicated nucleotide positions. For example, the

first fragment sequence is the part of the input sequence
between positions 10 and 62. Step 1308 then checks each
third subsequence probe subsequence against each fragment
sequence to determine whether there is any match (*i.e.*,
5 whether the probe has a sequence complementary enough to the
fragment sequence sufficient for it to hybridize thereon).
If a match is found, an indication is made in the fragment 4
member tuple. This match is done by string searching in a
similar manner to that described for generation of the end
10 vectors.

Next at step 1309 of Figure 16, all the fragment
are sorted on length and assembled into a vector of sorted
fragments, which is output from the mock fragmentation method
at step 1310. This vector contains the complete list of all
15 fragments, with probe information, defined by their end
subsequences and lengths that the input reaction will
generate from the input sequence.

Figure 17E illustrates the fragment vector of the
example sorted according to length. For illustrative
20 purposes, third subsequence probe P1 was found to hybridize
only to the third fragment 1412, where a 'Y' is marked. 'N'
is marked in all the other fragments, indicating no probe
binding.

The simulated database is generated by iteratively
25 applying the basic mock fragmentation method for each
sequence in the selected database and each reaction in the
experimental definition. Figure 18 illustrates a simulated
database generation method. The method starts at 1501 and at
1502 inputs the selected representative database and the
30 experimental definition with, in particular, the list of
reactions and their related subsequences. Step 1503
initializes the digest database table so that lists of
accession numbers may be inserted for all possible
combinations of fragment length and target end subsequences.
35 Step 1504, a DO loop, causes the iterative execution of steps
1505, 1506, and 1507 for all sequences in the input selected
database.

Step 1505 takes the next sequence in the database, as selected by the enclosing DO loop, and the next reaction of the experiment and performs the mock fragmentation method of Figure 16, on these inputs. Step 1506 adds the sorted

5 fragment vector to the simulated database by taking each fragment from the vector and adding the sequence accession number to the list in the database entry indexed by the fragment length and end subsequences and probe (if any). Figure 17F represents the simulated database entry list

10 additions that would result for the example mock fragmentation reaction of Figures 17A-E. For example, accession number A01 is added to the accession number list in the entry 1412 at length 151 and with both end subsequences RE2.

15   Finally, step 1507 tests whether there is another reaction in the input experiment that should be simulated against this sequence. If so, step 1505 is repeated with this reaction. If not, the DO loop is repeated to select another database sequence. If all the database sequences

20 have been selected, the step 1508 outputs the simulated database and the method ends at 1509.

5.4.5.2. QEA™ EXPERIMENTAL DESIGN METHODS

The goal of the experimental design methods is to

25 optimize each experiment in order to obtain the maximum amount of quantitative information. An experiment is defined by its component recognition reactions, which are in turn defined by the target end subsequences recognized, third subsequences recognized, if any, and labels assigned. If the

30 SEQ-QEA™ method or alternative phasing primers are used, effective target subsequences are used. Any of several criteria can be used to ascertain the amount of information obtained, and any of several algorithms can be used to perform the reaction optimization.

35   A preferred criteria for ascertaining the amount of information uses the concept of "good sequence." A good sequence for an experiment is a sequence for which there is

at least one reaction in the experiment that produces a
unique signal from that sequence, that is, a fragment is
produced from that good sequence, by at least one recognition
reaction, that has a unique combination of length and
5 labeling. For example, returning to Figure 15B, the sequence
with accession number A01 is a good sequence because reaction
1 produces signal 1215, with length 52 and with both target
end subsequences recognized by RE1, uniquely from sequence
A01. However, sequence S003 is not a good sequence because
10 there are no unique signals produced only from S003:
reaction R2 produces signal 1216 from both A01 and S003 and
signal 1219 from both Q012 and S003. Using the amount of
good sequences as an information measure, the greater the
number of good sequences in an experiment the better is the
15 experimental design. Ideally, all possible sequences in a
sample would be good sequences.

Further, a quantitative measure of the expression
of a good sequence can simply be determined from the detected
signal intensity of the fragment uniquely produced from the
20 good sequence. Relative quantitative measures of the
expression of different good sequences can be obtained by
comparing the relative intensities of the signal uniquely
produced from the good sequences. An absolute quantitative
measure of the expression of a good sequence can be obtained
25 by including a concentration standard in the original sample.
Such a standard for a particular experiment can consist of
several different good sequences known not to occur in the
original sample and which are introduced at known
concentrations. For example, exogenous good sequence 1 is
30 added at a $1:10^3$ concentration in molar terms; exogenous good
sequence 2 at a $1:10^4$ in molar terms, etc. Then comparison of
the relative intensity of the unique signal of a good
sequence in the sample with the intensities of the unique
signal of the standards allows determination of the molar
35 concentrations of the sample sequence. For example, if the
good sequence has a unique signal intensity half way between
the unique signal intensities of good sequences 1 and 2, then

- 183 -

it is present at a concentration half way between the
concentrations of good sequences 1 and 2.

Another preferred measure for ascertaining the
amount of information produced by an experiment is derived by
5 limiting attention to a particular set of sequences of
interest, for example a set of known oncogenes or a set of
receptors known or expected to be present in a particular
tissue sample. An experiment is designed according to this
measure to maximize the number of sequences of interest that
10 are good sequences. Whether other sequences possibly present
in the sample are good sequences is not considered. These
other sequences are of interest only to the extent that the
sequences of interest produce uniquely labeled fragments
without any contribution from these other sequences.

15   The QEA™ method experimental design is adaptable to
other measures for ascertaining information from an
experiment. For example, another measure is to minimize on
average the number of sequences contributing to each detected
signal. A further measure is, for example, to minimize for
20 each possible sequence the number of other sequences that
occur in common in the same signals. In that case each
sequence is linked by common occurrences in fragment
labelings to a minimum number of other sequences. This can
simplify making unambiguous signal peaks of interest (see
25 infra).

Having chosen an information measure, for example
the number of good sequences, for an experiment, the
optimization methods choose target subsequences, and possibly
probes, which optimize the chosen measure. One possible
30 optimization method is exhaustive search, in which all
subsequences in lengths less than approximately 10 are tested
in all combinations for that combination which is optimum.
This method requires considerable computing power, and the
upper bound is determined by the computational facilities
35 available and the average probability of occurrence of
subsequences of a given length. With adequate resources, it
is preferable to search all sequences down to a probability

of occurrence of about 0.005 to 0.01. Upper bounds may range
from 8 to 11 or 12.

A preferred optimization method is known as
simulated annealing. See Press et al., 1986, *Numerical*
5 *Recipes - The Art of Scientific Computing*, § 10.9, Cambridge
University Press, Cambridge, U.K. Simulated annealing
attempts to find the minimum of an "energy" function of the
"state" of a system by generating small changes in the state
and accepting such changes according to a probabilistic
10 factor to create a "better" new state. While the method
progresses, a simulated "temperature", on which the
probabilistic factor depends and which limits acceptance of
new states of higher energy, is slowly lowered.

In the application to the methods of the QEA™
15 method experimental design, a "state", denoted by S, is the
experimental definition, that is the target end subsequences
and hybridization probes, if any, in each recognition
reaction of the experiment. The "energy", denoted E, is
taken to be 1.0 divided by the information measure, so that
20 when the energy is minimized, the information is maximized.
Alternatively, the energy can be any monotonically decreasing
function of the information measure. The computation of the
energy is denoted by applying the function E( ) to a state.

The preferred method of generating a new
25 experiment, or state, from an existing experiment, or state,
is to make the following changes, also called moves to the
experimental definition: (1) randomly change a target end
subsequence in a randomly chosen recognition reaction; (2)
add a randomly chosen target end subsequence to a randomly
30 chosen reaction; (3) remove a randomly chosen target end
subsequence from a randomly chosen reaction with three or
more target subsequences; (4) add a new reaction with two
randomly chosen target end subsequences; and (5) remove a
randomly chosen reaction. All target end subsequences are to
35 be chosen from available RE recognition sequences. If the
SEQ-QEA™ method or alternative phasing primers are used to
generate effective target subsequences, all subsequences must

be chosen from among such effective target subsequences that can be generated from available REs. In the case of the SEQ-QEA™ method, the extra subsequence information is not known until the QEA™ method experiment is performed. To generate a

5 new experimental definition, one of these moves is randomly selected and carried out on the existing experimental definition. Alternatively, the various moves can be unequally weighted. In particular, if the number of reactions is to be fixed, moves (4) and (5) are skipped. The

10 QEA™ method is further adaptable to other moves for generating new experiments. Preferable generation methods will generate all possible experiments.

Several additional subsidiary choices are needed in order to apply simulated annealing. The "Boltzman constant"

15 is taken to be 1.0, so that the energy equals the temperature. The minimum of the energy and temperature, denoted E_0 and T_0, respectively, are defined by the maximum of the information measure. For example, if the number of good sequences of interest is G and is used as the information

20 measure, then E_0, which equals T_0, equals 1/G. An initial temperature, denoted T_1, is preferably chosen to be 1. An initial experimental definition, or state, is chosen, either randomly or guided by prior knowledge of previous experimental optimizations. Finally, two execution

25 parameters are chosen. These parameters define the "annealing schedule", that is the manner in which the temperature is decreased during the execution of the simulated annealing method. They are the number of iterations in an epoch, denoted by N, which is preferably

30 taken to be 100 and the temperature decay factor, denoted by f, which is preferably taken to be 0.95. Both N and f may be systematically varied case-by-case to achieve a better optimization of the experiment definition with a lower energy and a higher information measure.

35 With choices for the information measure or energy function, the moves for generating new experiments, an initial state or experiment, and the execution parameters

- 186 -

made as above, the general application of simulated annealing
to optimize an experimental definition is illustrated in
Figure 20A. The information measure used in this description
is the number of good sequences of interest. Any information
5 measure, such as those previously described, may be used
alternately.

The method begins at step 1701. At step 1702 the
temperature is set to the initial temperature; the state to
the initial state or experimental definition; and the energy
10 is set to the energy of the initial state. At step 1703 the
temperature and energy are checked to determine whether
either is less than or equal to the minima for the
information measure chosen, as the result of either a
fortuitous initial choice or subsequent computation steps.
15 If the energy is less than or equal to the minimum energy, no
further optimization is possible, and the final experimental
definition and its energy is output. If the temperature is
less than or equal to the minimum temperature, the
optimization is stopped. Then the inverse of the energy is
20 the number of good sequences of interest for this
experimental definition.

Step 1706 is a DO loop which executes an epoch, or
N iterations, of the simulated annealing algorithm. Each
iteration consists of steps 1707 through 1711. Step 1707
25 generates a new experimental definition, or state, S_{new},
according to the described generation moves. Step 1708
ascertains or determines the information content, or energy,
of S_{new}. Step 1709 tests the energy of the new state, and, if
it is lower than the energy of the current state, at step
30 1711, the new state and new energy are accepted and replace
the current state and current energy. If the energy of the
new state is higher than the energy of the current state,
step 1710 computes the following function.

$$EXP[-(E-E_{new})/T] \qquad\qquad (4)$$

35

This function defines the probabilistic factor controlling
acceptance. If this function is less than a random chosen
number uniformly distributed between 0 and 1, then the new
state is accepted at step 1711. If not, then the newly
5 generated state is discarded. These steps are equivalent to
accepting a new state if the energy is not increased by an
amount greater than that determined by function (4) in
conjunction with the selection of a random number. Or in
other words, a new state is accepted if the new information
10 measure is not decreased by an amount greater than indirectly
determined by function (4).

Finally, after an epoch of the algorithm, at step
1712 the temperature is reduced by the multiplicative factor
f and the method loops back to the test at step 1703.

15 Using this algorithm, starting from an initial
experimental definition which has certain information
content, the algorithm produces a final experimental
definition with a higher information content, or lower
energy, by repetitively and randomly altering the
20 experimental definition in order to search for a definition
with a higher information content.

The computation of the energy of an experimental
definition, or state, in step 1708 is illustrated more detail
in Figure 20B. This method starts at step 1720. Step 1721
25 inputs the current experimental definition. Step 1722
determines a complete digest database from this definition
and a particular selected database by the method of Figure
18. Step 1723 scans the entire digest database and counts
the number of good sequences of interest. If the total
30 number of good sequences is the measure used, the total
number of good sequences can be counted. Alternatively,
other information measures may be applied to the digest
database. Step 1724 computes the energy as the inverse of
the information measure. Alternatively, another decreasing
35 function of the information content may be used as the
energy. Step 1725 outputs the energy, and the method ends at
step 1726.

5.4.5.3. THE QEA™ METHOD AMBIGUITY RESOLUTION

In one utilization of the QEA™ method, DNA from two related tissue samples can be subject to the same experiment, perhaps consisting of only one recognition reaction, and the
5 outcomes compared. The two tissue samples may be otherwise identical except for one being normal and the other diseased, perhaps by infection or a proliferative process, such as hyperplasia or cancer. One or more signals may be detected in one sample and not in the other sample. Such signals might
10 represent genetic aspects of the pathological process in one tissue. These signals are of particular interest.

The candidate sequences that can produce a signal of interest are determined, as previously described, by look-up in the digest database. The signal may be produced by
15 only one sequence, in which case it is unambiguously identified. However, even if the experiment has been optimized, the signal may be ambiguous in that it may be produced by several candidate sequences from the selected database. A signal of interest may be made unambiguous in
20 several manners which are described herein.

In a first manner of making unambiguous assume the signal of interest is produced by several candidate sequences all of which are good sequences for the particular experiment. Then which sequences are present in the signal
25 of interest can be ascertained by determining the quantitative presence of the good sequences from their unique signals. For example, referring to Figure 15B, if the signal 1217 of length 175 with the labeling 1213 is of interest, the sequences actually present in the signal can be determined
30 from the quantitative determination of the presence of signals 1215 and 1218. Here, both the possible sequences contributing to this signal are good sequences for this experiment.

The first manner of making unambiguous can be
35 extended to the case where one of the sequences possibly contributing to a signal is not a good sequence. The quantitative presence of all the possible good sequences can

be determined from the quantitative strength of their unique
signals. The presence of the remaining sequence which is not
a good sequences can be determined by subtracting from the
quantitative presence of the signal of interest the
5 quantitative presences of all the good sequences.

Further extensions of the first manner can be made
to cases where more than one of the possible sequences is not
a good sequences if the sequences which are not good appear
as contributors to further signals involving good sequences
10 in a manner which allows their quantitative presences to be
determined. For example, suppose signal 1219 is of interest,
where both possible sequences are not good sequences. The
quantitative presence of sequence Q012 can be determined from
signals 1220 and 1218 in the manner previously outlined. The
15 quantitative presence of sequence S003 can be determined from
signals 1216 and 1215. Thereby, the sequences contributing
to signal 1219 can be determined. More complex combinations
can be similarly made unambiguous.

An alternative extension of the first manner of
20 making unambiguous is by designing a further experiment in
which the possible sequences contributing to a signal of
interest are good sequences even if they were not originally
so. Since there are approximately 50 suitable REs that can
be used in the RE embodiment of the QEA™ method (Section
25 6.2), there are approximately 600 RE reaction pairs that can
be performed, assuming that half of the theoretical maximum
of 1,250 (50 X 50 / 2 = 1,250) are not useable. Since most
RE pairs produce on the average of 200 fragments and standard
electrophoretic techniques can resolve at least approximately
30 500 fragment lengths per lane, the RE QEA™ method embodiment
has the potential of generating over 100,000 signals (500 X
200 = 100,000). The number of possible signals is further
increased by the use of reactions with three or more REs and
by the recognition of third subsequences. Further, since the
35 average complex human tissue, for example brain, is estimated
to express no more than approximately 25,000 genes, there is
a 4 fold excess of possible signals over the number of

- 190 -

possible sequences in a sample. Thus it is highly likely
that for any signal of interest, a further experiment can be
designed and optimized for which all possible candidates of
the signal of interest are good sequences. This design can
5 be made by using the prior optimization methods with an
information measure the sequences of interest in the signal
of interest and starting with an extensive initial
experimental definition including many additional reactions.
In that manner, any signal of interest can be made
10 unambiguous.

A second manner of making unambiguous is by
automatically ranking the likelihood that the sequences
possibly present in a signal of interest are actually present
using information from the remainder of the experimental
15 reactions. Figure 21 illustrates a preferred ranking method.
The method begins at step 1801 and at step 1802 inputs the
list of possible accession numbers in a signal of interest,
the experimental definition, and the actual experimental
results. DO-loop 1803 iterates once for each possible
20 accession number. Step 1804 performs a simulated experiment
by the method illustrated in Figure 11 in which, however,
only the current accession number is acted on. The output is
a single sequence digest table, such as illustrated in Figure
17F.

25   Step 1805 determines a numerical score of ranking
the similarity of this digest table to the experimental
results. One possible scoring metric comprises scanning the
digest table for all fragment signals and adding 1 to the
score if such a signal appears also in the experimental
30 results and subtracting 1 from the score if such signal does
not appear in the experimental results. Alternate scoring
metrics are possible. For example, the subtraction of 1 may
be omitted.

Step 1806 sorts the numerical scores of the
35 likelihood that each possible accession number is actually
present in the sample. Step 1807 outputs the sorted list and
the method ends at step 1808.

By this method likelihood estimates of the presence
of the various possible sequences in a signal of interest can
be determined.

5   5.4.6. <u>APPARATUS FOR PERFORMING THE QEA™ METHODS</u>
An apparatus for the QEA™ method includes means for
performing the computer implemented QEA™ experimental
analysis and design methods and optionally for performing the
QEA™ method recognition reactions in a preferably automated
10 fashion, for example by the protocols of § 6.1.12.1 (entitled
"QEA™ Preferred RE Method"). In the embodiment herein
presented both elements are described. In an alternative
embodiment, the laboratory methods can be performed by other
means, for example manually, and the apparatus needed is
15 limited to the computer apparatus described for performing
the experimental design and analysis methods.
Figure 19A illustrates an exemplary apparatus for
the QEA™ method embodiments. Computer 1601 can be,
alternatively, a UNIX based work station type computer, an
20 MS-DOS or Windows based personal computer, a Macintosh
personal computer, or another equivalent computer. In a
preferred embodiment, computer 1601 is a PowerPC™ based
Macintosh computer with software systems capable of running
both Macintosh and MS-DOS/Windows programs.
25   Figure 19B illustrates the general software
structure in RAM memory 1650 of computer 1601 in a preferred
embodiment. At the lowest software level is Macintosh
operating system 1655. This system contains features 1656
and 1657 for permitting execution of UNIX programs and MS-DOS
30 or Windows programs alongside Macintosh programs in computer
1601. At the next higher software level are the preferred
languages in which the QEA™ computer methods are implemented.
LabView 1658, from National Instruments (Dallas, TX), is
preferred for implementing control routines 1661 for the
35 laboratory instruments, exemplified by 1651 and 1652, which
perform the recognition reactions and fragment separation and
detection. C or C++ languages 1659 are preferred for

- 192 -

implementing experimental routines 1662, which are described in § ? (entitled "QEA™ Analysis And Design Methods"). Less preferred but useful for rapid prototyping are various scripting languages known in the art. PowerBuilder 1660,
5 from Sybase (Denver, CO), is preferred for implementing the user interfaces to the computer implemented routines and methods. Finally, at the highest software level are the programs implementing the described computer methods. These programs are divided into instrument control routines 1661
10 and experimental analysis and design routines 1662. Control routines 1661 interact with laboratory instruments, exemplified by 1651 and 1652, which physically perform the QEA™ method and CC protocols. Experimental routines 1662 interact with storage devices, exemplified by devices 1654
15 and 1653, which store DNA sequence databases and experimental results.

Returning to Figure 19A, although only one processor is illustrated, alternatively, the computer methods and instrument control interface can be performed on a
20 multiprocessor or on several separate but linked processors, such that instrument control methods 1661, computational experimental methods 1661, and the graphical interface methods can be on different processors in any combination or sub-combination.
25   Input/output devices include color display device 1620 controlled by a keyboard and standard mouse 1603 for output display of instrument control information and experimental results and input of user requests and commands. Input and output data are preferably stored on disk devices
30 such as 1604, 1605, 1624, and 1625 connected to computer 1601 through links 1606. The data can be stored on any combination of disk devices as is convenient. Thereby, links 1606 can be either local attachments, whereby all the disks can be in the computer cabinet(s), LAN attachments, whereby
35 the data can be on other local server computers, or remote links, whereby the data can be on distant servers.

Instruments 1630 and 1631 exemplify laboratory
devices for performing, in a partly or wholly automatic
manner, the QEA™ method recognition reactions. These
instruments can be, for example, automatic thermal cyclers,
5 laboratory robots, and controllable separation and detection
apparatus, such as is found in the applicants' copending U.S.
Patent Application 08/438,231 filed May 9, 1995, incorporated
by reference herein in its entirety. Links 1632 exemplify
control and data links between computer 1601 and controlled
10 devices 1631 and 1632. They can be special buses, standard
LANs, or any suitable link known in the art. These links can
alternatively be computer readable medium or even manual
input exchanged between the instruments and computer 1601.
Outline arrows 1634 and 1635 exemplify the physical flow of
15 samples through the apparatus for performing experiments 1607
and 1613. Sample flow can be either automatic, manual, or
any combination as appropriate. In alternative embodiments
there may be fewer or more laboratory devices, as dictated by
the current state of the laboratory automation art.
20   On this complete apparatus, a QEA™ method
experiment is designed, performed, and analyzed, preferably
in a manner as automatic as possible. First, a QEA™ method
experiment is designed, according to the methods specified in
§ 5.4.5 (entitled "QEA™ Analysis And Design Methods") as
25 implemented by experimental routines 1662 on computer 1601.
Input to the design routines are databases of DNA sequences,
which are typically representative selected database 1605
obtained by selection from input comprehensive sequence
database 1604, as described in § 5.4.5 (entitled "QEA™
30 Analysis And Design Methods"). Alternatively, comprehensive
DNA databases 1604 can be used as input. Database 1604 can
be local to or remote from computer 1601. Database selection
performed by processor 1601 executing the described methods
generates one or more representative selected databases 1605.
35 Output from the experimental design methods are tables,
exemplified by 1609 and 1615, which, for a QEA™ method RE

embodiment, specify the recognition reaction and the REs used
for each recognition reaction.

Second, the apparatus optionally performs the
designed experiment. Exemplary experiment 1607 is defined by
5 tissue sample 1608, which may be normal or diseased,
experimental definition 1609, and physical recognition
reactions 1610 as defined by 1609. Where instrument 1630 is
a laboratory robot for automating reaction, computer 1601
commands and controls robot 1630 to perform reactions 1610 on
10 cDNA samples prepared from tissue 1608. Where instrument
1631 is a separation and detection instrument, the results of
these reactions are then transferred, automatically or
manually, to 1631 for separation and detection. Computer
1601 commands and controls performance of the separation and
15 receives detection information. The detection information is
input to computer 1601 over links 1632 and is stored on
storage device 1624, along with the experimental design
tables and information on the tissue sample source for
processing. Since this experiment uses, for example,
20 fluorescent labels, detection results are stored as
fluorescent traces 1611.

Experiment 1613 is processed similarly along sample
pathway 1633, with robot 1630 performing recognition
reactions 1616 on cDNA from tissue 1608 as defined by
25 definition 1615, and device 1631 performing fragment
separation and detection. Fragment detection data is input
by computer 1601 and stored on storage device 1625. In this
case, for example, silver staining is used, and detection
data is image 1617 of the stained bands.
30   During experimental performance, instrument control
routines 1661 provide the detailed control signals needed by
instruments 1630 and 1631. These routines also allow
operator monitoring and control by displaying the progress of
the experiment in process, instrument status, instrument
35 exceptions or malfunctions, and such other data that can be
of use to a laboratory operator.

Third, interactive experimental analysis is
performed using the database of simulated signals generated
by analysis and design routines 1662 as described in § 5.4.5
(entitled "QEA™ Analysis And Design Methods"). Simulated
5 database 1612 for experiment 1607 is generated by the
analysis methods executing on processor 1601 using as input
the appropriate selected database 1605 and experimental
definition 1609, and is output in table 1612. Similarly
table 1618 is the corresponding simulated database of signals
10 for experiment 1613, and is generated from appropriate
selected database 1605 and experimental definition 1615. A
signal is made unambiguous by experimental routines 1662 that
implement the methods described in § 5.4.5 (entitled "QEA™
Analysis And Design Methods").
15   Display device 1602 presents an exemplary user
interface for the QEA™ method data. This user interface is
programmed preferably by using the Powerbuilder display front
end. At 1620 are selection buttons which can be used to
select the particular experiment and the particular reaction
20 of the experiment whose results are to be displayed. Once
the experiment is selected, histological images of the tissue
source of the sample are presented for selection and display
in window 1621. These images are typically observed,
digitized, and stored on computer 1601 as part of sample
25 preparation. The results of the selected reaction of the
selected experiment are displayed in window 1622. Here, a
fluorescent trace output of a particular labeling is made
available. Window 1622 is indexed by marks 1626 representing
the possible locations of DNA fragments of successive integer
30 lengths.
          Window 1623 displays contents from simulated
database 1612. Using, for example, mouse 1603, a particular
fragment length index 1626 is selected. The processor then
retrieves from the simulated database the list of accession
35 numbers that could generate a peak of that length with the
displayed end labeling. This window can also contain further
information about these sequences, such as gene name,

bibliographic data, etc. This further information may be available in selected databases 1605 or may require queries to the complete sequence database 1604 based on the accession numbers. In this manner, a user can interactively inquire

5 into the possible sequences causing particular results and can then scan to other reactions of the experiment by using buttons 1620 to seek other evidence of the presence of these sequences.

It is apparent that this interactive interface has

10 further alternative embodiments specialized for classes of users of differing interests and goals. For a user interested in determining tissue gene expression, in one alternative, a particular accession number is selected from window 1623 with mouse 1603, and processor 1601 scans the

15 simulated database for all other fragment lengths and their recognition reactions that could be produced by this accession number. In a further window, these lengths and reactions are displayed, and the user allowed to select further reactions for display in order to confirm or refute

20 the presence of this accession number in the tissue sample. If one of these other fragments are generated uniquely by this sequence (a "good sequence", see *supra*), that fragment can be highlighted as of particular interest. By displaying the results of the generating reaction of that unique

25 fragment, a user can quickly and unambiguously determine whether or not that particular accession number is actually present in the sample.

In another interface alternative, the system displays two experiments side by side, displaying two

30 histological images 1621 and two experimental results 1622. This allows the user to determine by inspection signals present in one sample and not present in the other. If the two samples were diseased and normal specimens of the same tissue, such signals would be of considerable interest as

35 perhaps reflecting differences due to the pathological process. Having a signal of interest, preferably repeatable and reproducible, a user can then determine the likely

accession numbers causing it by invoking the previously
described interface facilities. In a further elaboration of
this embodiment, system 1601 can aid the determination of
signals of interest by automating the visual comparison by
5 performing statistical analysis of signals from samples of
the same tissue in different states. First, signals
reproducibly present in tissue samples in the same state are
determined, and second, differences in these reproducible
signals across samples from the several states are compared.
10 Display 1602 then shows which reproducible signals vary
across the states, thereby guiding the user in the selection
of signals of interest.

     This apparatus has been described above in an
embodiment adapted to a single site implementation, where the
15 various devices are substantially local to computer 1601 of
Figure 19A, although the various links shown could also
represent remote attachments. Alternative, explicitly
distributed embodiments of this apparatus are possible as is
apparent to those of ordinary skill in the computer arts.

20   All the computer implemented QEA™ methods can be
recorded for storage and transport on any computer readable
memory devices known in the art. For example, these include,
but are not limited to, semiconductor memories - such as
ROMs, PROMs, EPROMs, EEPROMS, etc., of whatever technology or
25 configuration - magnetic memories - such as tapes, cards,
disks, etc of whatever density or size - optical memories -
such as optical read-only memories, CD-ROM, or optical
writable memories - and any other computer readable memory
technologies.
30

6. EXAMPLES

     The following examples further illustrate the
different features of the invention but do not in any way
limit the scope of the invention which is defined by the
35 appended claims. This section describing examples has been
divided into a section describing protocols that are common

to several of the examples and another section that is a
description of the examples themselves.

6.1.   DESCRIPTION OF PROTOCOLS
5   The following sections describe protocols for use.

6.1.1.   MATING PROTOCOL
          Mating of the yeast a and α strains is preferably
performed according to a filter disc mating protocol, which
10 achieves efficient cell handling, limited cell doublings, and
high mating efficiencies. An alternate less preferred
protocol is a plate mating protocol, which has less favorable
characteristics. After mating according to either protocol,
the mating efficiency preferably is estimated according to a
15 protocol which determines the ratio of the number of yeast
diploids to the total number of yeast cells.
          The filter disc protocol is preferred since more
cells can be mated with high mating efficiencies and with
fewer cell doublings during mating than can be achieved by
20 prior protocols, in particular by the plate mating protocol.
According to filter disc mating, at least approximately
3×10^3 cells, to at least approximately 6×10^0 cells, to
approximately 1×10^9 cells, to approximately 2×10^9 cells,
and up to approximately 3.5×10^9 cells can be mated per 90 mm
25 filter disc. (These cell numbers correspond approximately to
mating cell densities of approximately at least 5×10^4, at
least 1×10^9, at least 1.5×10^5, at least 3.5×10^5, and up
to $4-6 \times 10^6$ cells, respectively, per square millimeter on the
filter-disc during mating.) In contrast, plate mating is
30 limited to, for example, approximately 1×10^8 mating cells on
each 105 mm plate (a mating cell density of 6×10^3 cells per
square millimeter).
          Therefore, the filter disc mating is more efficient
in that it uses fewer mating resources, and consequently is
35 capable of processing mating experiments of greater
complexity, which require a greater number of mated cells.
Further, according to filter disc mating, no more than

approximately one cell doubling occurs during the conditions
prevailing during the mating period, whereas with plate
mating several cell doublings can occur during mating. Thus,
interacting colonies observed after filter disc mating are
5 more likely to represent independent and unique protein-
protein interactions than are the colonies observed after
cells mating. Finally, both filter disc and plate mating
usually achieve similarly high mating efficiencies (fraction
of diploids formed) of approximately 25% to 50%. This
10 invention is, however, adaptable to other mating protocols
that achieve efficient cell handling, limited cell doublings
during mating, and comparable mating efficiencies.

In summary, according to the filter disc mating
protocol, transformed yeast cells are grown to from mid to
15 late log phase to stationary phase on media selective for the
appropriate transforming plasmids, and then are briefly
boosted on rich medium immediately prior to mating. The
boosted cells of both mating strains are mixed in numbers
sufficient according to the statistical considerations
20 disclosed in Section 5.2.7. Aliquots of the mixed cells are
packed by, e.g., vacuum suction onto filter discs, which can
be of paper, nylon, or any other suitable material capable of
retaining yeast cells. The filter discs with the packed
cells are incubated at a temperature and for a time
25 sufficient to allow cell mating. Finally, mated cells are
harvested and transferred to media selective for appropriate
for diploids. Optionally, an aliquot of the harvested cells
is used to estimate the mating efficiency.

In more detail, a preferred embodiment of this
30 protocol proceeds according to the following detailed steps.
First, prior to mating, yeast cells bearing activation and
binding domain fusion plasmid libraries are grown for at
least two days, or until stationary phase, on media selective
for the appropriate plasmid. Stationary phase cells are then
35 "boosted" just prior to mating by a brief growth period on
rich media to numbers 3 to 5 fold higher than required for
mating. A volume of 1-2 ml of stationary phase library yeast

- 200 -

is diluted in 1000 ml of YPAD media (Sherman et al., eds., 1991, *Getting started with yeast*, Vol. 194, Academic Press, New York) and grown for 4-8 hours at approximately 30° C. Where one of the libraries is of limited complexity, for

5 example of a complexity, for example of complexity less than 10 or less than 50, it is advantageous to maintain the library members in separate cultures and to separately boost each member for 4-8 hours in YPAD medium.

Next, the boosted cells are mixed to form the

10 mating mix. The number of cells from each of the binding and activation domain libraries to be mixed is preferably determined according to the statistical considerations of Section 5.2.7. Alternately, and equally preferably, the number of cells to be mixed can be simply determined

15 according to the relation $F*M*N$, where F is a factor, M is the complexity of the binding domain library, and N is the complexity of the activation domain library. "Library complexity" is taken herein to mean the number of separate clones in the library. The factor F is approximately at

20 least 50, more preferably 75, or even more preferably 100 or greater. Cell number can be found from measurement of OD_{600}, where 1 OD_{600} unit equals approximately 2 x 10^7 cell/ml. Where one of the libraries is of limited complexity and the library members are maintained in separate cultures, an equal number

25 of each library member is mixed to attain the required cell number.

Next, aliquots of cells from the mating mix are packed onto filter discs soaked in rich medium, preferably, by vacuum aspiration. When a preferred 90 mm diameter filter

30 disc is used, the aliquots contain preferably between 1.5 and 2.0 x 10^9 cells and more preferably approximately 1.8 x 10^9 cells. For filter discs of other diameters, the preferred number of cells can be scaled according to the relative areas of the discs. A sufficient number of filter discs is used to

35 accommodate the total number of cells in the mating mix. As soon as the cells are packed on the filter disc, vacuum aspiration is stopped and the filter disc is placed on a

large YPAD plate, taking care that no air bubbles remain
between the filter disc and the plate. The plate(s) carrying
the filter disc(s) is then incubated for approximately 6-10
hours at approximately 30° C to permit cell mating. A

5   preferred filter disc is Catalog no. HAWP 090 25 from the
Millipore Corporation (Bedford, MA), and has a diameter of 90
mm with a pore size of 0.45 μm. A preferred vacuum
aspiration unit is a 500 ml large filtration unit from the
Fisher Scientific Corporation (Pittsburgh, PA).

10   Finally, after the mating incubation, the mated
cells are suspended in 1000 ml of sterile water by swirling
the filter disc(s), and are then screened for protein-protein
interactions by plating on appropriate media selective for
diploid cells bearing interacting binding domain and

15   activating-domain fusion proteins. For screening efficiency
and effectiveness, it is advantageous to plate no more than
approximately 50-100 expected interactant colonies or no more
the approximately 10^8 expected diploid cells per plate. These
expected numbers can be simply obtained as follows. The

20   expected number of diploids can be simply found by
multiplying the density of mated cells and the mating
efficiency, where the cell density can be estimated from the
OD_{600} and the mating efficiency can be estimated according to
the following protocol. The expected number of interactants

25   among the mated cells can be found by further multiplying the
expected number of diploids by the expected rate of protein-
protein interactions. The latter rate can be estimated from
experience with various mating, and in particular, it has
been found for libraries of interest derived from human

30   samples that the expected rate of protein-protein
interactions is approximately 2-6 x 10^{-7}. Using these
expected numbers, one of skill in the art will be able to
plate the mated cells according to the preferred criteria.
Even with such careful plating, however, screens of complex

35   libraries, which require large numbers of mated cells, can
require many 10s or even a few hundred plates.

Briefly, mating efficiency can be estimated by
plating serial dilutions of an aliquot of the suspended,
mated cells. The OD_{600}, and thus the cell concentration, is
measured after resolving cell flocculation by adding EDTA up
5 to a concentration of 2 mM. Serial dilutions from 10^{-1} to 10^{-9}
are then plated onto each of three plates, a first plate
selective for activation domain plasmids, a second plate
selective for binding domain plasmids, and a third plate
selective for diploid cells. Mating efficiency is estimated
10 from the set of plates with easily counted colonies as twice
the ratio of the number of diploid colonies to the sum of the
number of colonies containing each of the plasmids. An
independent estimate of the cell density can also be obtained
from the serial dilution plates.

15   In addition to the filter disc protocol, mating is
performed as per standard protocols (Sherman et al., eds.,
1991, *Getting started with yeast*, Vol. 194, Academic Press,
New York). Briefly, for the plate mating protocol, cells are
grown until mid to late log phase on solid or liquid media
20 that select for the appropriate plasmids. The two mating
strains, a and α, are then mixed together as a paste onto a
rich solid media like YPAD (Sherman et al., eds., 1991,
Getting started with yeast, Vol. 194, Academic Press, New
York) and incubated at 30°C for 6-8 hr. The cells are then
25 transferred to selective media appropriate for the desired
diploids.

In a preferred embodiment of the plate mating
protocol, 1×10^8 cells/ml of each mating type are mixed for
30 minutes at room temperature and then plated onto a 150 mm
30 diameter YPAD plate and incubated at 30°C for 6-8 hours.
Then, the contents of the plate are harvested in a volume of
1-2 ml in the appropriate selective media and transferred to
a 150 mm diameter plate that has the selective medium for
selecting interactions. Alternatively, the YPAD plate with
35 the mating mix can be replica-plated onto another 15 cm
diameter plate that has the selective medium for selecting
interactions.

- 203 -

6.1.2. TRANSFORMATION PROTOCOL

Yeast transformations are performed by the lithium
acetate procedure (Ito et al., 1983, J. Bacteriol.
153:163-168) and the transformants are selected by plating on
5 appropriate selective media that are usually Synthetic
Complete (SC) media that lack the appropriate nutrients
(Sherman et al., eds., 1991, *Getting started with yeast*, Vol.
194, Academic Press, New York).

In detail, lithium acetate transformation proceeds
10 according to the following steps. Cells to be transformed
are grown overnight in rich medium like YPAD medium (Sherman
et al., eds., 1991, *Getting started with yeast*, Vol. 194,
Academic Press, New York) and then diluted two-fold in rich
medium and shaken for two hours at 30° C. The cells are
15 pelleted and washed with sterile water and with
transformation buffer (0.1 M LiAc in 10 X TE buffer at pH
7.5). The washed cells are pelleted and resuspended in three
times the pellet volume of transformation buffer. In an
Eppendorf tube, to 80 μl of this cell suspension, are added
20 28 μg of single-stranded salmon sperm DNA in 10 X TE buffer,
and 1-10 μg of appropriate, transforming plasmid DNA, and
which is then incubated at room temperature for 5-10 minutes.
Then to each Eppendorf tube, are added 500 μl of a mixture of
40% PEG with a molecular weight of 3350 and 60%
25 transformation buffer, which are incubated at 30° C for 20-20
minutes, after which 58 μl of DMSO is added. The cells are
heat-shocked for 10-15 minutes in a 42-45° C water-bath,
washed in TE buffer, resuspended in 200 μl of TE buffer, and
plated onto appropriate selective media.
30

6.1.3. RNA EXTRACTION

The tissue to be extracted is weighed and a 10-fold
volume/weight of Triazol reagent (Life Technologies,
Gaithersburg, MD) is added and the tissue ground with a
35 Polytron homogenizer (Brinkman Instruments, Westbury, NY).
Example: 100 mg in 1 ml, 1 g in 10 ml. 0.2 volumes of
chloroform are added and vortexed for 15 seconds, and phases

separated by centrifugation (5000 x g, 15 min). The aqueous
phase is precipitated with 0.6 volumes of 2-propanol. The
precipitated RNA is pelleted at 10,000 x g for 15 min, rinsed
with 70% ethanol and dried. The RNA pellet is resuspended in
5 water to give a final concentration of 100 ng/μl.

6.1.4. DNASE TREATMENT

0.2 volumes of 5x reverse transcriptase buffer
(Life Technologies), 0.1 volumes of 0.1 M DTT, and 5 units
10 RNAguard/100 mg starting tissue (Pharmacia Biotech, Uppsala,
Sweden) are added to the RNA extracted according to Section
6.1.3. One unit RNase-free DNase I (Pharmacia Biotech)/100
mg starting tissue is added, and the mixture is incubated at
37°C for 20 min. 10 volumes of Triazol is added and RNA
15 extraction by addition of chloroform and precipitation is
repeated.

6.1.5. MESSENGER RNA PURIFICATION

RNA concentration is estimated by measuring OD_{260} of
20 a 100-fold dilution of extracted RNA mixture after DNase
treatment. The Dynal oligo(dT) magnetic beads have a
capacity of 1 μg poly(A+) per 100 μg of beads (1 mg/ml
concentration). Assuming that 2% of the total RNA is
poly(A+), 5 volumes of Lysis/Binding buffer (Dynal, Oslo,
25 Norway) and sufficient beads to bind poly(A+) are added.
This mixture is heated at 65°C for 2 min and then incubated
at room temperature for 5 min. The beads are first washed
with 1 ml washing buffer/LiDS (Dynal), then with 1 ml washing
buffer (Dynal, Oslo, Norway) twice. The poly(A+) RNA is
30 eluted with 1 μl water/μg beads twice.

6.1.6. cDNA SYNTHESIS AND CONSTRUCTION OF FUSION-LIBRARIES

cDNA synthesis is performed using the Hybrizap Two-
35 Hybrid cDNA synthesis and Gigapack cloning kit cDNA synthesis
kit (Stratagene) according to the manufacturer's protocol
with the following modifications. The cDNA synthesis is

- 205 -

performed substantially as per the Gubler-Hoffman method
(Gubler and Hoffman, 1983, Gene 25:263-269). In the first
strand synthesis step, MoMuLV reverse transcriptase is used
for reverse transcription. The primer (from the kit)
5 (GAGAGAGAGAGAGAGAGAGAACTAGTCTCGAGTTTTTTTTTTTTTTTTTTT)
(SEQ ID NO:36) used in the first strand synthesis also adds
an XhoI site near the 3' end. After the second strand
synthesis, EcoRI adapters (also from the kit) are ligated to
the cDNAs using standard linker ligation conditions according
10 to manufacturer's (Stratagene's) protocols. The identities
of the EcoRI adapters are AATTCGGCACGAG (SEQ ID NO:37) and
CTCGTGCCG (SEQ ID NO:38). Following this, the cDNA is
digested with EcoRI and XhoI and cloned into the EcoRI and
XhoI sites of the Hybrizap vector (Stratagene), which is a
15 lambda phage vector, using the manufacturer's protocols. The
phagemid pAD-GAL4 bearing the cDNA inserts is removed by *in
vivo* excision using the reagents and protocols provided in
the Hybrizap Gigapack cloning kit. This creates a cDNA
library, containing plasmid pAD-GAL4, with the sense strand
20 being in frame with the GAL4 activation domain of the plasmid
pAD-GAL4 (Stratagene). Plasmid pAD-GAL4 contains *LEU2* to
facilitate selection in media lacking leucine.

         In a different embodiment, the activation domain
fusion library is created in the vector pACT2 (Clontech).
25 The EcoRI-XhoI linked cDNA is cloned between the EcoRI and
SalI sites in pACT2. This creates a cDNA library with the
sense strand being in frame with the GAL4 activation domain
in the plasmid pACT2 (Clontech). Plasmid pACT2 contains *LEU2*
to facilitate selection in media lacking leucine.
30   In the case of cloning into pAS2-1 (Clontech) or
pBD-GAL (Stratagene) to create a library of DNA-binding
domain fusion genes, the EcoRI-XhoI linked cDNA is cloned
between the EcoRI and SalI sites in pAS2-1 or pBD-GAL to
create a cDNA library in plasmid pAS2-1 or pBD-GAL, with the
35 sense strand being in frame with the GAL4 DNA-binding domain.
Statistically, one in every three clones will represent a

true open reading frame. Plasmids pAS2-1 or pBD-GAL contain
TRP1 to facilitate selection in media lacking tryptophan.

6.1.7. TRANSFORMATION OF THE REPORTER STRAINS WITH THE BINDING DOMAIN FUSION cDNA LIBRARY AND ACTIVATION DOMAIN cDNA LIBRARY TO CREATE "M" AND "N" POPULATIONS

The strains YULH and N106' (see Sections 6.3.2 and
6.3.4) are transformed with the pAS2-1, pBD-GAL, and the pAD-
GAL4 or pACT2 cDNA libraries, respectively, by lithium
acetate protocol (Section 6.1.2; Ito et al., 1983, J.
Bacteriol. 153:163-168). One μg of library DNA generally
yields a maximum of 1 x 10^6 transformants. The transformants
are selected on either media lacking leucine (for pAD-
GAL4/pACT2) or lacking tryptophan and containing 5-FOA (for
pAS2-1 or pBS-GAL). In the latter case, all GAL4 DNA-binding
domain (GBD)-fusions that fortuitously activate transcription
on their own will be eliminated since 5-FOA kills the URA+
cells. It is preferred that 5-FOA negative-selection be
performed according to the protocol to be subsequently
described. The transformants are harvested in the
appropriate media (SC-Leu for pAD-GAL4/pACT2 and SC-TRP for
pAS2-1 or pBD-GAL) to a final cell density of 2 x 10^8 to 2 x
10^{10} cells/ml and preferably 2 x 10^9 cells/ml and stored in
aliquots at -70°C after making them 10% in DMSO or glycerol.

Negative selection of the binding domain library
transformants to eliminate fortuitous activation of the
reporter genes is, as has been described, always important
but is especially so in the case of complex activation or
binding domain libraries. Since fortuitous activation can
occur in up to 1-5% of binding domain transformants, without
such negative selection, finding the occasional protein-
protein interaction among the numerous false-positive,
fortuitously activating binding domain transformants is
virtually impossible. For example, a binding domain library
of complexity 10^7 with a fortuitous activation rate of 1%
results in approximately 10^5 false positive colonies for each
activation domain library member. Individually screening

- 207 -

such a vast number of false-positive colonies for true
protein-protein interactions is clearly quite impractical.
Effective use of complex libraries depends on negative
screening protocols which greatly reduce fortuitously
5 activating binding domain transformants.

Since it has been found that fortuitous activation
by activating-domain fusions with the GAL4 activating domains
are almost never observed, negative-selection of activating-
domain transformants is not usually useful.

10 In more detail, the preferred negative selection
protocol achieves a fortuitous activation rate of preferably
less than approximately 5×10^{-6}, or less than approximately
4×10^{-6}, or less than approximately 3×10^{-6}, or less than
approximately 2×10^{-6}, or preferably less than approximately
15 1×10^{-6}, or even less. Simple plating of binding domain
libraries on plates that negatively select for the expression
of reporter genes such as URA3, LYS2, Can1, or CYH2 has been
found to result in a fortuitous activation rate of no less
than approximately 10^{-4} to approximately 10^{-5} in the harvested
20 cells. However, most advantageously, where URA3 is used as
one of the reporter genes, it has been found that negative
selection with 5-FOA according to the following protocol has
been observed to routinely reduce fortuitous activation to a
rate of less than approximately 1×10^{-6}. If a fortuitous
25 activation rate greater than approximately 1×10^{-6} is found,
further protocol steps replica plating (as described below)
are performed. Accordingly, this embodiment is most
preferred for binding domain libraries of any complexity, and
especially for complex binding domain libraries.

30 The preferred 5-FOA negative selection protocol
proceeds according to the following steps. Approximately
2×10^{8} cells transformed with the binding domain library are
shaken for approximately 2 hours at 30°C, pelleted, and then
resuspended in 50 ml of sterile water. Using the cell
35 density calculated from the measured OD_{600} (1 OD_{600} unit equals
approximately 2×10^{7} cells/ml), an aliquot containing
approximately 10^{5} cells is plated on a large plate containing

- 208 -

media selective for the binding domain plasmid and containing 5-FOA. A sufficient number of plates is plated so that the total number of cells plated equals approximately three times the complexity of the binding domain library.

5   After overnight incubation at room temperature, the plate(s) is incubated at 30°C until the colonies grow up. These colonies are then replica plated to another large plate with the same medium lacking tryptophan and containing 5-FOA, and once again the colonies are allowed to grow up. After 10 2-3 days, the colonies are harvested by scraping and pooling, and the cells are stored in 15 glycerol and 3% DMSO.

The replica plating step is important in achieving the extra reduction in fortuitous activation rate. Optionally, this replica plating step can be repeated until 15 the fortuitous activation rate no longer declines. The fortuitous activation rate at each step of replica plating can be estimated by plating serial dilutions of a sample of harvested cells on medium selective for the reporter gene, and finding the ratio of positive colonies to the total cells 20 plated known from the cell density. It has been found that a single replica plating achieves most of the decrease in fortuitous activation, and that subsequent replica platings generally do not result in further significant decreases. Replica plating is the preferred method of selectively 25 removing only yeast cells that are actively growing in the toxic environment from substantially all other yeast cells, including dead cells, cells which are living but not viable, and cells which are dormant in the toxic environment but still viable and capable of future growth in a non-toxic 30 environment. Further, any dormant URA3· cells that are transferred into a new media will enter into a new growth phase and will, thereby, be inhibited or killed by the 5-FOA.

Negative selection can also be done according to a bait validation protocol, which screens both fortuitously- 35 activating binding domain fusion proteins and also fusion proteins in the activation domain library or the binding domain-library that non-specifically associate with other

proteins, and thereby activate reporter gene expression.
Bait validation is most advantageously applied to matings in
which one library has such limited complexity (the "bait"
library) that each member can be separately manipulated and
5 separately maintained in individual cultures. Briefly, bait
validation separately mates each member of the bait library
with the more complex library and selects out and removes
from further consideration those bait-library members that
too frequently activate reporter gene expression. As
10 described in Section 5.2.8, for mammalian or human samples,
it is most preferred to select out those library members that
activate reporter gene(s) with a frequency greater than
approximately 10^{-3}.

In more detail, the more complex library is grown
15 for 4-8 hours by inoculating 1-2 ml of frozen library stock
(or enough stock to achieve an OD_{600} of approximately 0.2) in
500 ml of a rich medium like YPAD. After this growth, the
cell density is measured, for example from OD_{600} values, and
aliquots of approximately 50,000 colonies per plate are
20 plated on plates selective for the appropriate library
plasmid. Beginning on the second day of complex library
growth on these plates, each member of the low complexity
library is grown to stationary phase in media selective for
the appropriate library plasmid, and then 300 μl aliquots of
25 this stationary-phase culture are plated onto YPAD mating
plates. Then, the more complex library is also replica
plated onto the mating plates, which are then incubated for
10 hours and 30° C for cell mating to occur.

The mated cells are screened by replica plating
30 them onto two plates, one with media appropriately selective
for diploid cells and the other with media appropriately
selective for diploid cells with reporter gene activation.
Each member of the bait library for which the most preferred
rate of reporter gene activation is exceeded is not used
35 further.

6.1.8. INTERACTANT PCR

After mating (Section 6.1.1), PCR can be performed using cells positive for protein-protein interactions in order to discover the fusion fragments responsible for the interaction. PCR is preferably performed on DNA templates derived from lysed yeast cells in a 96 (or greater) well format. Less preferably, PCR is performed on whole cells, which are lysed at the denaturation temperature of the first PCR thermal cycle.

The preferable PCR protocols proceed, first, by producing yeast DNA template, and second, by PCR amplification of this template. Yeast DNA template is produced by treating an aliquot of yeast cells positive for interaction, first, with a cell-wall lytic enzyme, such as Zymolase, to dissolve cell walls, and second, with a proteolytic enzyme, such as Proteinase K, to inactivate all other lytic enzymes. Proteinase K self inactivates and need not be separately inactivated.

In detail, 10 μl of Zymolase solution is added to each well of a 384 or a 96 well PCR plate. Zymolase solution is 2.4 M sorbitol, 100 mM sodium phosphate buffer at pH 7.4, 60 mM β-mercaptoethanol, 1 mM EDTA, and 5 mg/ml of Zymolase. This solution is made by adding Zymolase to small aliquots of the sorbitol/sodium phosphate buffer just before use. An aliquot of 10 μl of yeast cells from colonies positive for protein-protein interaction (Section 6.1.9) is added to each well of the plate and the plate is incubated at room temperature for 30 minutes or at 37° C for 5 minutes. The samples are then held at 4° C until the next step. Next, 10 μl of 30 μg/ml of Proteinase K is added to each well, and the plate is incubated sequentially at 50° C for 10 minutes, 95° C for 10 minutes, and then held at 4° C.

Using this yeast DNA template product, PCR is preferably performed with a hot-start protocol. Hot-start protocols are advantageous to reduce false priming and primer-dimer formation. One preferred hot-start protocol proceeds by adding an essential PCR reaction component,

preferably the dNTPs, after the reaction mixture has reached the denaturation temperature of, for example, 94° C. A most preferred hot-start protocol proceeds by separating two components of the PCR reaction mix by a wax layer in a

5 reaction wells. The amplification only commences when the reaction mix has been sufficiently pre-heated to melt the wax layer and to allow the two components to mix.

A first preferred hot-start PCR reaction is done in a reaction volume of approximately 50 μl in wells of a 96

10 well microtiter plate. It will be apparent to those of skill in the art how to scale the reaction conditions for, e.g., 384 well microtiter plates. The following reactants are premixed and are added to each well:

41 μl	Water
5 μl	10 X PCR2 buffer (1 X PCR2 Buffer = 20 mm Tris-HCl pH 8.55, 16 mM ammonium sulfate, 2.5 mM $MgCl_2$, 150 μg/ml BSA)
0.2 μl	50 pm/μl of M13-40AD5 + BACREVAD3 (Ab Peptides, St. Louis, MO) for amplifying activation domain fusions
0.2 μl	50 pm/μl of pAS3BacREV + pASForM13-40 for amplifying binding domain fusions
0.3 μl	25 U/ml KlenTaq:Pfu (16:1 volume ratio)

15

20

Next add 1.5 μl of the appropriate yeast DNA template prepared according to the previous protocol to each well. Preferably, this contains approximately 1-10 ng of DNA. The microtiter plate is briefly equilibrated to 94° C for 15

25 seconds and 2 μl of 5 mM dNTPs are added to each well. The following thermal profile is then performed:

94° C	for 4 minutes after adding dNTPs;
94° C	for 40 seconds;
50° C	for 40 seconds;
72° C	for 3 minutes; then repeat 94-50-72° C for five cycles;
94° C	for 40 seconds;
58° C	for 40 seconds;
72° C	for 4 minutes; then repeat 94-58-72° C for 28 cycles;
72° C	for 5 minutes.

30

35

The PCR amplification is adaptable to certain variations of this thermal profile according to guidelines known in the

art.   For example, the reaction time at 72° C can be adjusted
for the expected length of products, generally allowing one
minute for each kilo-base.   A three minute time permits
amplification of up to three kilo-base fragments.   The cycle
5 numbers can be chosen according to the abundance of the yeast
template and the PCR reaction efficiency.   These numbers can
be sufficiently large to detect products but not so large
that amplification background interferes with product
detection.

10   The most preferred hot-start protocol is done in
pre-waxed 96-well PCR plates.   A preferred wax, which melts
at approximately 72° C is a 90:10 mixture of
Paraffin:Chillout™ 14.   The paraffin is a highly purified
paraffin wax melting between 58 °C and 60 °C such as can be
15 obtained from Fluka Chemical, Inc. (Ronkonkoma, N.Y.) as
Paraffin Wax cat. no. 76243.   Chillout™ 14 Liquid Wax is a
low melting, purified paraffin oil available from MJ
Research.   Pre-waxed PCR plates are made by layering
approximately 40 μl of the melted wax on the upper third of
20 the wall of each well in the PCR plate, and by allowing it to
solidify.   The PCR mix is divided into a "lower mix" and an
"upper mix," which individually do not react, of the
following compositions.

LOWER MIX:

25   25 μl   Water
            3 μl   10 X PCR2 buffer (1 X PCR2 Buffer = 20 mm
                       Tris-HCl pH 8.55, 16 mM ammonium sulfate,
                       2.5 mM MgCl$_2$, 150 μg/ml BSA)
            2 μl   dNTPs (5 nM equi-molar mixture))

UPPER MIX:

30   15.2 μl   Water
            2 μl   10 X PCR2 buffer
            0.25 μl   100 pm/μl of primer (M13-40AD5 for
                       activation domain fusions; pAS3BacREV for
                       binding domain fusions) (Ab Peptides, St.
                       Louis, MO)
            0.25 μl   100 pm/μl of primer (BACREVAD3 for
                       activation domain fusions; pASForM13-40
35   for binding domain fusions) (Ab Peptides,
                       St. Louis, MO)
            2 μl   5 M Betaine
            0.3 μl   25 U/ml KlenTaq:Pfu (16:1 volume ratio)

The protocol proceeds according to the following steps. 30 μl of the lower mix is dispensed into each PCR reaction well. Any droplets on the sides of the wells are centrifuged down for approximately 10 seconds. The wax is

5 then melted and solidified onto the top of the lower mix by carrying out the following thermal program: 72° C for 3 minutes; then 65° C, 55° C, and 50° C in turn for 1 minute each; then 45° C, 40° C, 35° C, 30° C in turn for 30 seconds each; then hold at 25° C. Next, 20 μl of the upper mix is

10 carefully added to each PCR well on top of the wax layer. Next, 2 μl of the appropriate yeast DNA template are added to each reaction well. PCR amplification is then performed according to the following thermal program:

94° C	for 4 minutes after adding dNTPs;

15
94° C	for 40 seconds;
50° C	for 40 seconds;
72° C	for 3 minutes; then repeat 94-50-72° C for five cycles;

| 94° C | for 40 seconds; |
| 58° C | for 40 seconds; |
20 | 72° C | for 4 minutes; then repeat 94-58-72° C for 28 cycles; |

| 72° C | for 5 minutes; |

| 4° C | hold. |

The reaction time at 72° C is chosen assuming that some of
25 the yeast DNA template will be up to 2 kb in size.

Advantageously, the fluid manipulation steps of this protocol can be performed by a standard laboratory robot, such as that available from the Tecan Corporation.

Finally, a less preferable, alternative, whole-cell
30 PCR is performed under the following conditions:

Reaction volume : 100 μl

10 X PC2 Buffer for Klentaq polymerase: 10 μl (1 X PC2 Buffer
=   20 mm Tris-HCl pH 8.55, 16 mM ammonium sulfate, 2.5 mM
      MgCl$_2$, 150 μg/ml BSA)
35 10 mM dNTPs : 3 μl

50 pmoles of each primer pair

1.0 µl of Klentaq polymerase (a thermostable DNA polymerase
        sold by AB Peptides Inc., St. Louis, MO).

2-5 µl of saturated culture of yeast in water.

PCR is performed at 94°C for 30 sec, 45-55°C for 30 sec and

5 72°C for 2 min, with each being repeated for 20-30 cycles.
The annealing temperature (i.e., the 45-55°C for 30 sec step)
depends on the melting temperature of the primers used. The
PCR primers are designed in such a way that the melting
temperature usually lies between 45-55°C.

10   A primer pair suitable for use according to either
PCR protocol can be selected from among those described
below. To amplify the fusion gene insert from pAS2, pAS2-1,
pASSfiI, pBD-GAL4, and other related vectors such as pAS1
(collectively referred to herein as "pAS-like vectors") (pAS1

15 is a parental GAL4-DNA binding domain vector; see Durfee et
al., 1993, Genes Dev. 7:555-569), one of the following primer
pairs can be used:

        pAS3BacREV + pASForM13-40
        pACTBAC + pASFOR
20   pASSEQI + pASSEQII
        pASSEQIA + pASSEQII

pASForM13-40, pASSEQI, and pASSEQIA are interchangeable.
pAS3BACREV and pACTBAC are interchangeable.

        To amplify the fusion gene insert from pACT, pACT2,
25 pACTSfiI, pAD-GAL4 and other related vectors (collectively
referred to herein as "pACT-like vectors"), one of the
following primer pairs can be used:

        M13-40 + BACREVAD3
        pACTBAC + pACTFOR
30   pACTBAC + pACTFORII
        pACTSEQI + pACTSEQII
        pACTSEQI + pACTBAC
        pACTSEQII + pACTFOR
        pACTSEQII + pACTFORII

35 BACREVAD3, pACTBAC and pACTSEQII are interchangeable.
M13-40AD5, pACTFORII, and pACTSEQI are interchangeable.

The identities of the above-listed primers are as follows:

pAS3BACREV =   5'-AGG AAA CAG CTA TGA CCA TCT GAG AAA GCA ACC
                TGA CCT (SEQ ID NO:118)

5 pASForM13-40 = 5'-GTT TTC CCA GTC ACG ACG GTG CGA CAT CAT CAT
                CGG AAG (SEQ ID NO:119)

M13-40AD5 =   5'-GTT TTC CCA GTC ACG ACG AGG GAT GTT TAA TAC
                CAC TAC (SEQ ID NO:120)

BACREVAD3 =   5'-AGG AAA CAG CTA TGA CCA TGC ACA GTT GAA GTG
10   AAC TTG C (SEQ ID NO:121)

pACTSEQII = 5'-CGA TGC ACA GTT GAA GTG AAC-3' (SEQ ID NO:1)

pACTFORII = 5'-CGC GTT TGG AAT CAC TAC AGG GAT G-3' (SEQ ID NO:2)

pACTBAC = 5'-CTA CCA GAA TTC GGC ATG CCG GTA GAG GTG TGG TCA-
15 3' (SEQ ID NO:3)

pASFOR = 5'-ATG AAG CTA CTG TCT TCT ATC GAA C-3'(SEQ ID NO:4)

pACTFOR = 5'-ATGGATGATGTATATAACTATCTATTC-3' (SEQ ID NO:122)

pACTSEQI = 5'-TTGGAATCACTACAGGGATG-3' (SEQ ID NO:49)

pASSEQI = 5'-GAATTCATGGCTTACCCATAC-3' (SEQ ID NO:50)

20 pASSEQII = 5'-AACCTGACCTACAGGAAAGAGTTAC-3' (SEQ ID NO:51)

pASSEQIA = 5'-CCTCTAACATTGAGACAGCATAG-3' (SEQ ID NO:52)

The primers can be used in sequencing as well as in PCR.

6.1.9. RECOVERY OF COLONIES POSITIVE FOR PROTEIN-PROTEIN INTERACTION

25   Colonies that are URA+, HIS+, and 3-AT^r are selected
as positive for protein-protein interactions and arrayed onto
96-well (or 384-well) plates in which each well contains
100 μl of the appropriate selective media like SC-URA-HIS-
30 TRP-LEU+3-AT (SC medium lacking uracil, histidine,
tryptophan, leucine, and containing 3-amino-1,2,4-triazole).
In an equally preferred mode, colonies that are URA+ and HIS+
are selected on plates lacking Tyr, Leu, Ura, His. Thus,
each well serves as source of a single colony positive for
35 protein-protein interactions, and each column or row in a 96-
well plate now serves as a pool of positive colonies. Cells
are grown at 30°C until late log phase (OD$_{600}$ of 1.5 - 2).

These cells are processed further or stored frozen at -80°C
after making them 10% in DMSO or glycerol.

Selection as above on plates with media entirely
deficient in products of the reporter genes may cause certain
5 weak protein-protein interactions to be missed. In certain
cases, it may be advantageous in order to detect such weak
protein-protein interactions to select on plates with trace
quantities of the reporter gene products. In particular, in
the case of the yeast strain YULH, the reporter gene URA3 can
10 have a low level of natural expression. Thereby, strong
protein-protein interactions are required for growth on media
entirely lacking in uracil. To detect weaker protein
interactions, it has been found advantageous to include a
trace amount of uracil in the selective media. It has been
15 found that adding approximately 1-10 μM, and preferably
approximately 5 μM, of uracil to the selective media allows
the detection of weak protein-protein interactions that would
otherwise have been missed.

20   6.1.10. PRODUCTION OF PCR POOLS FOR
                    CREATION OF PROTEIN INTERACTION MAPS
If the total number of positive colonies is less
than 1500 then they are readily pooled according to a two-
dimensional pooling scheme. 10 μl of each well in a given
25 column or row are combined into a single pool and mixed well.
The mix is centrifuged at 1000 g for 2 minutes, resuspended
in 100 μl of water, centrifuged again as described above, and
the supernatant discarded. The pelleted cells are preferably
lysed (Section 6.1.8), or less preferably, the PCR mix is
30 added directly to the pellet and mixed well. PCR is
performed wherein DNA-binding (pAS-specific or pBD-GAL
specific) and activation domain fusion specific primers (pAD-
GAL4/pACT-specific) amplify the genes encoding the two
interacting proteins directly from yeast (Section 6.1.8).
35 Thus, each PCR reaction refers to the "M" population or the
"N" population. Primers that can be used are described in
Section 6.1.8.

- 217 -

6.1.11.   β-GALACTOSIDASE ASSAYS

Filter-lift β-galactosidase assays are performed as
modified from the protocol of Breeden and coworkers (Breeden
and Nasmyth, 1985, Cold Spring Harb. Symp. Quant. Biol.
5 50:643-650). The URA+, HIS+ and 3-AT^r colonies are patched
onto SC-TRP-LEU-URA-HIS+3-AT plates, grown overnight and
replica plated onto Whatman no. 1 filter papers overlayed
onto SC-TRP-LEU plates and again grown overnight at 30°C.
The filters with the grown colonies of yeast are then assayed
10 for β-galactosidase activity. Colonies positive for
β-galactosidase activity turn blue. Quantitative
β-galactosidase assays on yeast are performed as described
previously by Coney and Roeder (Coney and Roeder, 1988, Mol.
Cell. Biol. 8:4009-4017). Chemiluminescent β-galactosidase
15 assays are performed by using the Galacto-Light and Galacto-
Light Plus Chemiluminescent reporter assay system for the
detection of β-galactosidase (Tropix, Inc.) according to the
manufacturer's protocols. Fluorescent β-galactosidase assays
are performed using the FluoReporter lacZ/Galactosidase
20 Quantitation kit (Molecular Probes) according to the
manufacturer's protocols.

In particular, a preferred protocol for performing
the filter-lift assay for β-galactosidase activity is
presented herein. An assay solution is prepared by combining
25 100 ml of Z-buffer, 0.27 ml β-mercaptoethanol, and 1 ml X-gal
stock (5-bromo-4-chloro-3-indolyl-d-D-galactoside at a
concentration of 33.4 mg/ml in N,N-dimethylformamide). (Z-
buffer is made by adding to 800 ml of water 16.1 g of Na_2HPO_4,
5.5 g of NaH_2PO_4, 0.75 g of KCl, and 0.246 g of $MgSO_4.7H_2O$,
30 adjusting the pH to 7.0, and adding water to 1000 ml.) For
smaller yeast growth plates, a 75 mm filter paper (Whatman 1
of VWR grade 413) is soaked in 1.8 ml of assay solution in a
petri dish. For larger growth plates, 3-4 ml of assay
solution is used with a correspondingly larger filter paper.
35 Yeast colonies are then lifted off the growth plate with
Optitran filter paper, Catalog no. BA-S 85 Schleicher and
Schull (Keene, N.H.), and the filter paper is placed with the

colonies facing up in a pool of liquid nitrogen for
approximately 5 seconds. Then the filter paper is thawed at
room temperature and then placed onto the filter paper soaked
with assay solution, taking care that no air bubbles remain
5 between the two filter papers. The filter papers are
incubated at 30-37° C for up to several hours. Positive β-
galactosidase activity is indicated by a blue color appearing
in from 1 minute to 10 hours.
zzzz

10   6.1.12. PROTOCOLS FOR QEA™ METHODS
                     AND SEQ-QEA™ METHODS

          6.1.12.1. PREFERRED QEA™ RE METHOD
          A DNA (preferably cDNA) population is input to the
QEA™ method protocols described in this section. This DNA
15 population can be pooled DNAs, each DNA encoding an
interactant protein identified according to the methods of
the invention, or can be, or can be derived from, one or both
of two DNA populations encoding the initial protein
populations between which (in fusion form) protein
20 interactions are detected according to the invention.
          This protocol is designed to keep the number of
individual manipulations down, and thereby raise the
reproducibility of the QEA™ method procedure. In a preferred
method, no buffer changes, precipitations or organic
25 (phenol/chloroform) extractions are used, all of which lower
the overall efficiency of the process and reduce its utility
for general use and more specifically for its use in
automated or robotic procedures.
          The protocol is described in terms of cDNA, but can
30 be used with any DNA.

          6.1.12.1.1. cDNA PREPARATION
          Terminal phosphate removal from cDNA is illustrated
with the use of Barents sea shrimp alkaline phosphatase
35 ("SAP") (U.S. Biochemical Corp.) and 2.5 μg of cDNA.
Substantially less (<10 ng) or more (>20 μg) of cDNA can be
prepared at a time with proportionally adjusted amounts of

enzymes. Volumes are maintained to preserve ease of
handling. The quantities necessary are consistent with
using the method to analyze small tissue samples from normal
or diseased specimens.

5   1.   Mix the following reagents
              2.5 μl   200 mM Tris-HCL
              23 μl   cDNA
              2 μl   2 units/μl Shrimp alkaline phosphatase
        The final resulting cDNA concentration is 100 ng/μl.
10   2.   Incubate at 37°C for 1 hour
        3.   Incubate at 80°C 15 minutes to inactivate the SAP.

6.1.12.1.2.   PREFERRED RE/LIGASE AND AMPLIFICATION REACTIONS
              Once the cDNA has been prepared, including terminal
15 phosphate removal, it is separated into a number of batches
of from 10 ng to 200 ng each, equal to the desired number of
individual samples that need to be analyzed and the extent of
the analysis. For example, if six RE/ligase reactions and
six analyses are needed to generate all necessary signals,
20 six batches are made. Shown by example are 50 ng fractions.
              RE/ligase reactions are performed as digestions by,
preferably, a pair of REs; alternatively, one or three or
more REs can be used provided the four base pair overhangs
generated by each RE differ and can each be ligated to a
25 uniquely adapter and a sufficiently resolved length
distribution results. The amount of RE enzyme specified is
sufficient for complete digestion while minimizing any other exo-
or endo-nuclease activity that may be present in the enzyme.
              Adapters are chosen that are unique to each RE in a
30 reaction. Thus, one uses a linker complementary to each
unique RE sticky overhang and a primer which uniquely
hybridized with that linker. The primer/linker combination
is an adapter, which will preferably be uniquely and
distinguishably labeled.
35

Adapter Annealing

Pairs of 12-mer linkers and 24-mer primers are pre-annealed to form adapters before they are used in the QEA™ method reactions, as follows:

5   1.   Add to water linker and primer in a 2:1 concentration ratio (12-mer : 24-mer) with the primer at a total concentration of 5 pM per μl.

2.   Incubate at 50°C for 10 minutes.

3.   Cool slowly to room temperature and store at -20°C.

10

Restriction-Digestion/Ligation Reaction

Reactions are prepared for use in a 96 well thermal cycler.   Add per reaction:

1.   1 U of appropriate REs (New England Biolabs,

15   Beverly, MA) (preferred RE pair listing in §

6.1.12.3 (entitled "Preferred QEA™ Method Adapters

and RE Pairs"))

2.   1 μl of appropriate annealed adapter

3.   1 μl of Ligase/ATP (0.2 μl T4 DNA ligase [1

20   U/μl]/0.8 μl 10 mM ATP from Life Technologies

(Gaithersburg, MD))

4.   0.5 μl 50 mM $MgCl_2$

5.   10 ng of subject prepared cDNA

6.   1 μl 10X NEB2 buffer from New England Biolabs

25   (Beverly, MA)

7.   Water to bring total volume to 10 μl

Then perform the RE/ligation reaction by following the thermal profile in Figure 22A using a PTC-100 Thermal Cycler from MJ Research (Watertown, MA).

30

Amplification Reaction

Prepare the PCR reaction mix by combining:

1.   10 μl 5X E-Mg (300 mM Tris-Hcl pH 9.0, 75 mM

$(NH_4)_2SO_4$, no Mg ions))

35   2.   100 pm of appropriate fluorescently labeled 24-mer

primers

3.   1 μl 10 mM dNTP mix (Life Technologies, Gaithersburg, MD)

4.   2.5 U of 50:1 Taq polymerase (Life Technologies, Gaithersburg, MD) : Pfu polymerase (Stratagene, La Jolla, CA)

5.   Water to bring volume to 40 μl per PCR reaction

Then perform the following steps:

1.   Add 40 μl of the PCR reaction mix to each RE/ligation reaction

2.   Perform the PCR temperature profile of Figure 22B using a PTC-100 thermal cycler (MJ Research, Watertown, MA)

6.1.12.1.3.   PREFERRED AUTOMATED RE/LIGASE REACTIONS

The reactions of the preceding section can be automated according to the following protocol which requires intermediate reagent additions or by a protocol note requiring such additions.

Single Tube Protocol With Reagent Additions

Reactions are preformed in a standard 96 well thermal cycler format using a Beckman Biomek 2000 robot (Beckman, Sunnyvale, CA). Typically 4 cDNA samples are analyzed in duplicate with 12 different RE pairs, for a total of 96 reactions. All steps are performed by the robot, including solution mixing, from user provided stock reagents, and temperature profile control.

Pre-annealed adapters are prepared as in the preceding section.

Restriction-Digestion/Ligation Reaction

Mix per reaction:

1.   1 U of appropriate RE (New England Biolabs, Beverly, MA)

2.   1 μl of appropriate annealed adapter (10 pmoles)

3.   0.1 μl T4 DNA ligase [1 U/μl] (Life Technologies (Gaithersburg, MD)

- 222 -

4. 1 μl ATP (Life Technologies, Gaithersburg, MD)

5. 5 ng of subject prepared cDNA

6. 1.5 μl 10X NEB2 buffer from New England Biolabs
    (Beverly, MA)

5   7. 0.5 μl of 50 mM $MgCl_2$

8. Water to bring total volume to 10 μl and transfer
    to thermal cycler

The robot requires 23 minutes total time to set up
the reactions. Then it performs the RE/ligation reaction by
10 following the temperature profile of Figure 22C using a PTC-
100 Thermal Cycler equipped with a mechanized lid from MJ
Research (Watertown, MA).

Amplification Reaction

15   Prepare the PCR reaction mix by combining:

1. 10 μl 5X E-Mg (300 mM Tris-HCl pH 9.0, 75 mM
    $(NH_4)_2SO_4)$

2. 100 pm of appropriate fluorescently labeled 24-mer
    primer

20 3. 1 μl 10 mM dNTP mix (Life Technologies,
    Gaithersburg, MD)

4. 2.5 U of 50:1 Taq polymerase (Life Technologies,
    Gaithersburg, MD) : Pfu polymerase (Stratagene, La
    Jolla, CA)

25 5. Water to being volume to 35 μl per PCR reaction
           Preheat the PCR mix to 72°C and transfer 35 μl of
the PCR mix to each digestion/ligation reaction and mix. The
robot requires 6 minutes for the transfer and mixing.

           Then the robot performs the PCR amplification
30 reaction by following the temperature profile of Figure 22B
using a PTC-100 thermal cycler equipped with a mechanized lid
(MJ Research, Watertown, MA).

           The total elapsed time for the digestion/ligation
and PCR amplification reactions is 179 minutes. No user
35 intervention is required after initial experimental design
and reagent positioning.

- 223 -

Single Tube Protocol Without Reagent Additions

First, add the PCR reaction mix by combining in the reaction tube:

1. 10 μl 5X E-Mg (300 mM Tris-HCl pH 9.0, 75 mM $(NH_4)_2SO_4$)

2. 100 pm of appropriate fluorescently labeled 24-mer primer

3. 2 μl 10 mM dNTP mix (Life Technologies, Gaithersburg, MD)

4. 2.5 U of 50:1 Taq polymerase (Life Technologies, Gaithersburg, MD) : Pfu polymerase (Stratagene, La Jolla, CA)

5. Water to bring volume to 40 μl per PCR reaction

Second, add a bead of wax melting approximately at 72°C (Ampliwax, Perkin-Elmer, Norwalk, CT). Melt the wax at 75°C for 5 minutes, and let the wax solidify at 25°C for 10 minutes with the lid open.

Third, add the RE/ligase reaction mix by combining in the reaction tube:

1. 0.1 μl of the REs (New England Biolabs, Beverly, MA)

2. 1 μl of appropriate annealed adapter (2:1 of 12:24 mer at 50 pmoles/ml)

3. 0.2 μl T4 DNA ligase [1 U/μl] (Life Technologies (Gaithersburg, MD)

4. 1 μl of 0.1 M ATP (Life Technologies, Gaithersburg, MD)

5. 1 μl of subject prepared cDNA (0.1-10 ng)

6. 0.1 μl 10X NEB 2 buffer from New England Biolabs (Beverly, MA)

7. 0.5 μl of 50 mM $MgCl_2$

8. Water to bring total volume to 10 μl and transfer to thermal cycler

Then perform the RE/ligation and PCR reactions by following the thermal profile in Figure 22D using, for example, a PTC-100 Thermal Cycler from MJ Research (Watertown, MA).

6.1.12.1.4. ALTERNATIVE RE/LIGASE AND AMPLIFICATION REACTIONS

Once the cDNA has been prepared it is separated
into a number of batches of from 20 ng to 200 ng each equal
5 to the desired number of individual samples that need to be
analyzed and the extent of the analysis. For example, if six
RE/ligase reactions and six analyses are needed to generate
all necessary signals, six batches are made. Shown by
example are 50 ng fractions.

10   RE/ligase reactions are performed as digestions by,
preferably, a pair of REs; alternatively, one or three or
more REs can be used provided the four base pair overhangs
generated by each RE differ and can each be ligated to a
uniquely adapter and a sufficiently resolved length
15 distribution results. The amount of RE enzyme specified is
sufficient for complete digestion while minimizing any other exo-
or endo-nuclease activity that may be present in the enzyme.

RE Digestion

20   Digest (with 50 ng of cDNA)

       1.   Mix the following reagents

           0.5 μl   prepared cDNA (100 ng/μl) mixture

           10 μl   New England Biolabs Buffer No. 2

           3 Units   RE enzyme

25   2.   Incubate for 2 hours at 37°C. Larger size digests
           with higher concentrations of cDNA can be used and
           fractions of the digest saved for additional sets
           of experiments.

Adapter Ligation

30   Since it is important to remove unwanted ligation
products, such as concatamers of fragments from different
cDNAs resulting from hybridization of RE sticky ends, the
restriction enzyme is left active during ligation. This
35 leads to a continuing cutting of unwanted concatamers and end
ligation of the desired end adapters.

The majority of restriction enzymes are active at the 16°C ligation temperature. Ligation profiles consisting of optimum ligation conditions interspersed with optimum digestion conditions can also be used to increase efficiency
5 of this process. An exemplary profile comprises periodically cycling between 37°C and 10°C and 16°C at a ramp of 1°C/min.

One linker complementary to each 5 minutes overhang generated by each RE is required. 100 pico moles ("pm") is a sufficient molar excess for the protocol described. For each
10 linker a complementary uniquely labeled primer is added for ligation to the cut ends of cDNAs. 100 pm is a sufficient molar excess for the protocol described. If the amounts of RE cDNA is changed the linker and primer amounts should be proportionately changed.
15

Ligation Reaction

(per 10 μl and 50 ng cDNA)

1. Mix the following reagents

Component	Volume
20	RE digested cDNA mixture	10 μl
100 pM/μl each primer	1 μl
100 pM/μl each linker	1 μl

2. Thermally cycle from 50°C to 10°C (-1°C/minute) then back to 16°C

25 3. Add 2 μl 10 mM ATP with 0.2 μl T4 DNA ligase (Premix 0.1 μl ligase 1 U/μl per 1 μl ATP) (E. Coli ligase is a less preferred alternative ligase.)

4. Incubate 12 hours at 16°C. This step can be shortened to less than 2 hours with proportionately
30   higher ligase concentration. Alternately the thermal cycling protocol described can be used here.

5. Incubate 2 hours 37°C

6. Incubate 20 minutes at 65°C to heat inactivate the
35   ligase (last step should be RE cutting).

7. Hold at 4°C

Amplification Of Fragments With Ligated Adapters

This step amplifies the fragments that have been cut twice and ligated with adapters unique for each RE cut end. It is designed for a very high amplification

5 specificity. Multiple amplifications are performed, with an increasing number of amplification cycles. Use the minimum number of cycles to get the desired signal. Amplifications above 20 cycles are not generally reliably quantitative.

Mix the following to form the ligation mix:

Component	Volume
RE/Ligase cDNA mixture	5 μl
10X PCR Buffer	5 μl
25 mM MgCl$_2$	3 μl
10 mM dNTPs	1 μl
100 pM/μl each primer	1 μl

Mix the following to form 150 μl PCR-Premix

30 μl	Buffer E (ligation mix will contribute 0.3 mM MgCl)
1 μl	(300 pmoles/μl Rbuni24 Flour) 24 mer primer strand (50 pmoles/μl NBuni24 Tamra)
0.6 μl	Taq polymerase (per 150 μl)
3 μl	dNTP (10 mM)
106 μl	H$_2$O

25   Amplification of fragments is more specific if the small linker dissociates from the ligated primer-cDNA complex prior to amplification. The following is an exemplary method for amplification of the results of six RE/ligase reactions.

1.   Place three strips of six PCR tubes, marked 10, 15,
30   and 20 cycles, into three rows on ice as shown.

```
20 cycles    1   2   3   4   5   6- Add 140 μl PCR-premix
15 cycles    1   2   3   4   5   6
10 cycles    1   2   3   4   5   6- Add 10 μl ligation mix
```

35   2.   Place 10 μl ligation mix in each tube in 10 cycle row

3.   Place 140 μl PCR premix in each tube in 20 cycle row

4.   Place into cycler and incubate for 5 minutes at 72°C. This melts linker which was not covalently ligated to the second strand of a cDNA fragment and allows the PCR premix to come to temperature.

5.   Move the 140 μl PCR premix into the tubes in the 10 cycle row containing the 10 μl ligation mix, then place 50 μl of result into corresponding tubes each in other rows.

6.   Incubate for 5 minutes at 72°C. This finishes incompletely double stranded cDNA ends into complete dsDNA, the top primer being used as template for second strand completion.

The amplification cycle is designed to raise specificity and reproducibility of the reaction. High temperature and long melting times are used to reduce bias of amplification due to high G+C content. Long extension times are used to reduce bias in favor of smaller fragments.

7.   Thermally cycle 95°C for 1 minute followed by 68°C for 3 minutes. Long denaturing times reduce PCR bias due to melting rates of fragments, and long extension time reduces PCR bias on fragment sizes.

8.   Incubate at 72°C for 10 minutes at end of reaction.

6.1.12.1.5. OPTIONAL POST-AMPLIFICATION STEPS

Several optional steps can improve the signal from the detected bands. First, single strands produced as a result of linear amplification from singly cut fragments can be removed by the use of single strand specific exonuclease. Exo I is the preferred nuclease.

1.   Incubate 2 units of nuclease with the product of each PCR reaction for 60 minutes at 37°C.

Second, the amplified products can be concentrated prior to detection either by ethanol precipitation or column separation with a hydroxyapatite column.

Several labeling methods are usable, including fluorescent labeling as has been described, silver staining, radiolabelled end primers, and intercalating dyes. Fluorescent end labeling is preferred for high throughput
5 analysis with silver staining preferred if the individual bands are to be removed from the gel for further processing, such as sequencing.

Finally, fourth, use of two primers allows direct sequencing of separated strands by standard techniques. Also
10 separated strands can be directly cloned into vectors for use in RNA assays such as in situ analysis. In that case, it is more preferred to use primers containing T7 or other polymerase signals.

15   6.1.12.2.   PREFERRED METHODS OF A SEQ-QEA™ EMBODIMENT
          6.1.12.2.1.   QEA™ METHOD PREFERRED FOR
                        USE IN A SEQ-QEA™ METHOD
The following single tube RE/ligase and PCR protocol is the most preferred embodiment of a QEA™ method,
20 not only when employing a SEQ-QEA™ method.

Initially 10 ng of each pooled PCR product (e.g., binding domain fusion proteins; activation domain fusion proteins) is digested with two restriction enzymes that each recognize a 4 nucleotide restriction site (like Sau3AI,
25 BsaWI, or Tsp509I). After that, the restriction enzymes are destroyed either by heat inactivation or by extraction with phenol and chloroform. The restriction digestion is done in a volume of 50 μl and the digested DNA is extracted and precipitated. The digested DNA is then used as input to a
30 QEA™ method reaction.
      Reagents Used:
      •   RE enzymes (RE1 and RE2)
      •   primer set 1 and primer set 2
      •   cDNA
35   •   10mM ATP
      •   10X NEB Buffer 2 (10mM Tris HCl pH 7.9, 10mM $MgCl_2$, 50mM NaCl, 1mM DTT (dithiothreitol))

- T4 DNA ligase
- 5 M betaine
- 10mM dNTP (equimolar mixture of all 4 dNTPs)
- 10X TB2.0 buffer (50mM Tris pH 9.15, 16mM $(NH_4)_2SO_4$, 2mM $MgCl_2$)
- 16 units Klentaq (Ab Peptides, Inc.): 1 unit Pfu polymerase (Stratagene, Inc.)
- wax(90:10 Parafin:Chillout PCR wax)
- water

A pair of RE enzymes, RE1 and RE2, to perform the QEA™ method are selected according to Sec. 6.1.12.3. For RE1 (or RE2), primer set 1 (or primer set 2) comprising a primer and a linker are also selected according to Sec. 6.1.12.3, specifically, Table 10.

The following components are mixed in a 1.5 ml tube to form QPCR mix, quantities as shown:

Reagent	1 rxn
TB 2.0	5 μl
dNTP	2 μl
Klentaq	0.25 μl
water	32.75 μl

The solutions are mixed by tapping and/or inverting the solution. Pre-waxed PCR tubes are used where 90:10 Paraffin:Chillout wax had been melted and added to the tubes in such a way that the wax solidified on the sides of the upper half of the tube. 40μl QPCR mix is added to the prewaxed PCR tubes, avoiding the sides and wax in the tubes. The tubes are placed in a thermal cycler without lids and the wax is melted onto the liquid layer by incubating at 75°C for 2 min, followed by decreasing increments of 5°C for every 2 min until 25°C is reached.

The following components are mixed as shown, to form the Qlig mix:

Reagent	1 rxn
Primer set 1	1μl
Primer set 2	1μl
ATP	0.8μl
NEB Buffer 2	1μl
Betaine	2μl
Digested DNA	1μl
T4 DNA Ligase	0.2μl
H_2O	3μl

The Q-lig mixes are added to the top of the wax layer in the
PCR tubes containing the Q-PCR mix. Caps are applied gently
to the PCR tubes and PCR is performed under the following
conditions: 37°C for 30 minutes, followed by a decrease to
16°C with a decrease of 1°C every minute. This is followed
by an incubation at 16°C for 1 hr, followed by an incubation
at 37°C for 10 minutes. This is followed by an incubation at
65°C for 10 minutes, followed by an incubation at 72°C for 20
minutes. After this, 20 cycles of the following conditions
are repeated: 96°C for 30 seconds, 57°C for 1 minute and
72°C for 2 minutes. This is followed by an incubation at
72°C for 10 minutes and then the QEA™ method reactions are
stored at -20°C until processed further.

QEA™ Method Post-Processing Protocol ("Biotin bead clean-up")

        In the case where one of the primers in either
primer set 1 or primer set 2 has an attached biotin capture
moiety at its 5' end, this post-processing protocol purifies
the QEA™ method reaction products and denatures the DNA
strands for analysis of the strand not captured via the
biotin moiety.

Reagents Used:

QEA™ method reaction samples

Dynal Magnetic Streptavidin Beads

Binding Buffer: 5M NaCl, 10 mM Tris, pH 8.0, 1 mM EDTA

Wash Buffer I: 10 mM Tris, pH 8.0

Wash Buffer II: 10 mM EDTA

Loading Buffer: Deionized formamide, 25 mM EDTA (pH 8.0), 50
mg/ml Blue dextran (1000 μl formamide is mixed with 200 μl
EDTA/dextran)

Ladder Loading Buffer: 100 μl GeneScan 500 ROX ladder
5 (molecular weight makers) (Applied Biosciences, Inc. (ABI),
Norwalk, CT) with 900 μl Loading Buffer

The magnetic streptavidin beads are washed with 3
volumes of binding buffer and then resuspended in an equal
volume of binding buffer. An equal volume of beads is added
10 to the QEA™ method reaction to be processed (e.g., 5 μl beads
for 5 μl QEA™ method sample). Purifications are done in a 96
well Costar PCR plate. The QEA™ method products are added to
the beads and incubated at room temperature for 15 minutes.
These are then placed on a Tecan magnetic holder and the
15 magnetic beads are allowed to migrate down. The supernatant
is discarded and washed with 200 μl of wash buffer I,
followed by a 200 μl wash with wash buffer II. When a
SEQ-QEA™ method is to be done, then the additional procedures
starting with digestion with Type IIS restriction enzymes
20 described in Section 6.1.12.2.2 are inserted here, prior to
air drying and resuspension in loading buffer. If a SEQ-QEA™
method is not to be done, the beads are then air-dried and
resuspended in loading buffer (5 μl for 5 μl of beads). In
the loading buffer the GeneScan 500 ROX ladder may be mixed
25 in a one-tenth dilution. The processed QEA™ method samples
are then analyzed by electrophoresis on an ABI 377 (Applied
Biosystems, Inc.) automated sequencer using the GeneScan
software (ABI) for analysis.

30   6.1.12.2.2. SEQ-QEA™ METHOD STEPS
When a SEQ-QEA™ method is to be done, the QEA™
method is carried out through the washing and purification
procedures involving wash buffer II of the biotin bead clean-
up, except that the QEA™ method primer pairs (primer set 1
35 and primer set 2) are replaced by SEQ-QEA™ method primer
pairs. One of these SEQ-QEA™ method primers has a Type IIS
restriction enzyme (e.g., Fok I) recognition site and a

- 232 -

fluorescent tag, (e.g., FAM (carboxy-fluoroscein) (ABI))
attached at the 5' end. The other primer has a biotin tag
("Bio") used for QEA™ method processing and comprises either
a uracil residue or a site for a rare-cutting restriction
5 enzyme like AscI. Sec. 6.1.12.5 and Table 18 has a list of
exemplary primers and linkers for the SEQ-QEA™ methods.

The following are preferable primers and linkers to
be used together with the REs BglII and BspHI.

SEQ-QEA™ method primer pairs	Type-IIS Enzyme	Method of Bead Release
1) KA5/KA24-FAM + RC9/UC24-Bio	FokI	UDG
2) BA5/BA24-FAM + RC9/UC24-Bio	BbvI	UDG
3) KA5/KA24-FAM + RC9/SC24-Bio	FokI	AscI
4) BA5/BA24-FAM + RC9/SC24-Bio	BbvI	AscI

15   Using the above REs and primer pairs, the QEA™ method
reaction products obtained fall into the following three
categories:

a)   A double-stranded DNA with a 5' FAM label with nearby
      sequence containing a recognition site for FokI or BbvI
20   on one strand, and a 3' biotin label with nearby
      sequence containing a uracil residue or an AscI
      recognition site on the other strand (in the case where
      different REs cut at each end)

b)   A double-stranded DNA with a 5' biotin label with nearby
25   sequence containing a uracil residue or an AscI
      recognition site on one strand, and a 3' biotin label
      with nearby sequence containing a uracil residue or an
      AscI recognition site on the other strand (in the case
      where same RE cuts at both ends)

30 c)   A double-stranded DNA with a 5' FAM label with nearby
      sequence containing a recognition site for FokI or BbvI
      on one strand, and a 3' FAM label with nearby sequence
      containing a recognition site for FokI or BbvI on the
      other strand (in the case where same RE cuts at both
35   ends)

After the biotin bead clean-up, that is, washing
and purification procedures using magnetic streptavidin beads

- 233 -

as described above through the use of wash buffer II, only
category "a" will be visible to fluorescent analysis.
Typically, after the reaction is completed, 45 µl out of 50
µl is processed (the rest is saved). These 45 µl of the QEA™
5 method reaction are bound to the magnetic streptavidin beads
as described above. Subsequently, the DNA bound to the beads
is digested with the Type IIS restriction enzyme in a volume
of 100 ml with about 10 units of the enzyme for 3 hours at
37°C. Type IIS restriction enzymes cleave DNA at a location
10 outside their recognition sites, thus producing overhangs of
unknown sequences (Szybalski et al., 1991, *Gene* 100:13-26).
The Type IIS digestion thus releases the FAM label and
creates a fragment-specific overhang that acts as a template
for sequencing. The supernatant is then removed and the
15 beads are washed with wash buffer I followed by a wash with
wash buffer II.

        The end-sequencing reaction is essentially a fill-
in reaction using the overhang generated by the Type-IIS
restriction enzyme as a template. Dideoxy chain terminators
20 labeled with different ABI fluorescent dyes are mixed at high
ratios with dNTPs to ensure high frequency of incorporation,
and the DNA polymerase enzyme used (*e.g.*, Sequenase (T7 DNA
polymerase), Taquenase (Taq polymerase)) has high affinity
for the labeled dideoxynucleotides. A sequencing mix
25 totalling 20 µl containing the appropriate 1x buffer, 1 µl
dNTPs diluted 1/200 from stock (3 mM dATP, 1.2 mM dCTP,
4.5 mM dGTP, 1.2mM dTTP), 0.5 µl each ABI dye-labeled
terminator solution (containing ddATP, ddCTP, ddGTP and
ddTTP, respectively), (and 1 µl 0.1 M DTT for Sequenase) is
30 made. The beads are resuspended in the sequencing mix and
0.1 µl Taquenase is added and the reaction is incubated at
65°C for 15 minutes. If Sequenase is to be used, 0.1 µl
Sequenase is added instead of taquenase and the reaction is
incubated at 37°C for 15 minutes. After this, the reaction
35 mix is transferred to a magnet and the supernatant is
removed. The beads are washed twice with wash buffer I.

The above-described end-sequencing reaction incorporates dye labeled nucleotides into the strand that contains biotin. Since biotin-streptavidin binding is nearly irreversible, the labeled strands must be cleaved for
5 analysis by electrophoresis. This is achieved by treating UMP-containing fragments with Uracil DNA Glycosylase (UDG), or cleaving AscI-site-containing fragments with Asc I. UDG removes the Uracil residue from dsDNA; the phosphate backbone is subsequently hydrolyzed at temperatures above room
10 temperature and at pH>8.3.

For UDG treatment, the beads are resuspended in 20 µl UDG buffer (30 mM Tris-HCl pH 7.5, 50 mM KCl, 5 mM MgCl$_2$), 0.2 units of UDG are added and the reaction is incubated at room temperature for 30 minutes. The reaction is then
15 transferred to a magnet and the supernatant removed. The biotinylated strand, which is the strand that is being filled in during end-sequencing, is still attached to the beads as UDG does not destroy the backbone, but makes it very susceptible to hydrolysis.

20   The beads are resuspended in 5 µl formamide loading buffer. These are then split into 2 tubes of 2.5 µl each. Another 2.5 µl formamide loading buffer is added to one and 2.5 µl formamide loading buffer with 20% GS500 ROX ladder (ABI) is added to the other. These are heated at 95°C for 5
25 minutes to effect hydrolysis and denaturation and analyzed by electrophoretic separation.

In case of the biotinylated primer having an Asc I site, the following is performed. The beads are resuspended in 20 µl of Asc I buffer and 5 units of Asc I is added and
30 incubated at 37°C for 1 hour. The beads are separated on a magnet and the supernatant that contains the digestion products is precipitated with three volumes of ethanol after the addition of 5 µg of glycogen. The pellet is resuspended in 5 µl formamide loading buffer and split into 2 tubes of
35 2.5 µl each. Another 2.5 µl formamide loading buffer is added to one and 2.5 µl formamide loading buffer with 20% GS500Rox ladder is added to the other. These are heated at

95°C for 5 minutes and analyzed by electrophoretic separation.

Sequencing is completed by gel electrophoretic separation of released and sequenced strands. The overhang
5 sequence is the order of partially filled in fragments observed.

6.1.12.3. PREFERRED QEA™ METHOD ADAPTERS AND RE PAIRS

Table 2 lists preferred primer-linker pairs that
10 may be used as adapters for the preferred RE embodiment of a QEA™ method.   The primers listed cover all possible double-digest RE combinations involving approximately 56 available RE having a 5' 4 bp overhang. There are 40 such REs available from New England Biolabs. For each QEA™ method
15 double digest, one primer and one linker from the "R" series and one primer and one linker from the "J" series are used together. This choice satisfies all adapter constraints previously described. Two pairs from the same series are not compatible during amplification.
20

TABLE 2:   SAMPLE ADAPTERS

Series	Adapter: Primer (longer strand) Linker (shorter strand)		RE
RA24	5' AGC ACT CTC CAG CCT CTC ACC GAA 3' (SEQ ID NO:53)		
RA1	(SEQ ID NO:54)	3' AG TGG CTT TTAA	Tsp509I Mfe1 EcoRI
RA5	(SEQ ID NO:55)	3' AG TGG CTT GTAC	NcoI BspHI
RA6	(SEQ ID NO:56)	3' AG TGG CTT GGCC	XmaI NgoMI BspEI
RA7	(SEQ ID NO:57)	3' AG TGG CTT GCGC	BssHII AscI
RA8	(SEQ ID NO:58)	3' AG TGG CTT GATC	AvrII NheI XbaI

RA9	(SEQ ID NO:59)	3' AG TGG CTT CTAG	DpnII BamHI BclI
RA10	(SEQ ID NO:60)	3' AG TGG CTT CGCG	KasI
RA11	(SEQ ID NO:61)	3' AG TGG CTT CCGG	EagI Bsp120I NotI EaeI
RA12	(SEQ ID NO:62)	3' AG TGG CTT CATG	BsiWI Acc65I BsrGI
RA14	(SEQ ID NO:63)	3' AG TGG CTT AGCT	XhoI SalI
RA15	(SEQ ID NO:64)	3' AG TGG CTT ACGT	ApaLI
RA16	(SEQ ID NO:65)	3' AG TGG CTT AATT	AflII
RA17	(SEQ ID NO:66)	3' AG TGG CTT AGCA	BssSI
RC24	5' AGC ACT CTC CAG CCT CTC ACC GAC 3' (SEQ ID NO:67)		
RC1	(SEQ ID NO:68)	3' AG TCG CTG TTAA	Tsp509I EcoRI ApoI
RC3	(SEQ ID NO:69)	3' AG TCG CTG TCGA	HindIII
RC5	(SEQ ID NO:70)	3' AG TCG CTG GTAC	BspHI
RC6	(SEQ ID NO:71)	3' AG TCG CTG GGCC	AgeI NgoMI BspEI SgrAI BsrFI BsaWI
RC7	(SEQ ID NO:72)	3' AG TCG CTG GCGC	MluI BssHII AscI
RC8	(SEQ ID NO:73)	3' AG TCG CTG GATC	SpeI NheI XbaI

RC9	(SEQ ID NO:74)	3' AG TCG CTG CTAG	DpnII BglII BamHI BclI BstYI SauIIIA
RC10	(SEQ ID NO:75)	3' AG TCG CTG CGCG	KasI
RC11	(SEQ ID NO:76)	3' AG TCG CTG CCGG	Bsp120I NotI
RC12	(SEQ ID NO:77)	3' AG TCG CTG CATG	Acc56I BsrGI
RC14	(SEQ ID NO:78)	3' AG TCG CTG AGCT	SalI
RC15	(SEQ ID NO:79)	3' AG TCG CTG ACGT	Ppu10I ApaLI
JA24	5' ACC GAC GTC GAC TAT CCA TGA AGA 3' (SEQ ID NO:80)		
JA1	(SEQ ID NO:81)	3' GT ACT TCT TTAA	Tsp509I Mfe1 EcoRI
JA5	(SEQ ID NO:82)	3' GT ACT TCT GTAC	NcoI BspHI
JA6	(SEQ ID NO:83)	3' GT ACT TCT GGCC	XmaI NgoMI BspEI
JA7	(SEQ ID NO:84)	3' GT ACT TCT GCGC	BssHII AscI
JA8	(SEQ ID NO:85)	3' GT ACT TCT GATC	AvrII NheI XbaI
JA9	(SEQ ID NO:86)	3' GT ACT TCT CTAG	DpnII BamHI BclI
JA10	(SEQ ID NO:87)	3' GT ACT TCT CGCG	KasI
JA11	(SEQ ID NO:88)	3' GT ACT TCT CCGG	EagI Bsp120I NotI EaeI

JA12		3' GT ACT TCT CATG	BsiWI
	(SEQ ID NO:89)		Acc65I BsrGI
JA14		3' GT ACT TCT AGCT	XhoI
	(SEQ ID NO:90)		SalI
JA15		3' GT ACT TCT ACGT	ApaLI
	(SEQ ID NO:91)		
JA16		3' GT ACT TCT AATT	AflII
	(SEQ ID NO:92)		
JA17		3' GT ACT TCT AGCA	BssSI
	(SEQ ID NO:93)		
JC24	5' ACC GAC GTC GAC TAT CCA TGA AGC 3' (SEQ ID NO:94)		
JC1		3' GT ACT TCG TTAA	Tsp509I
	(SEQ ID NO:95)		EcoRI ApoI
JC3		3' GT ACT TCG TCGA	HindIII
	(SEQ ID NO:96)		
JC5		3' GT ACT TCG GTAC	BspHI
	(SEQ ID NO:97)		
JC6		3' GT ACT TCG GGCC	AgeI
	(SEQ ID NO:98)		NgoMI BspEI SgrAI BsrFI BsaWI
JC7		3' GT ACT TCG GCGC	MluI
	(SEQ ID NO:99)		BssHII AscI
JC8		3' GT ACT TCG GTAC	SpeI
	(SEQ ID NO:100)		NheI XbaI
JC9		3' GT ACT TCG CTAG	DpnII
	(SEQ ID NO:101)		BglII BamHI BclI BstYI
JC10		3' GT ACT TCG CGCG	KasI
	(SEQ ID NO:102)		
JC11		3' GT ACT TCG CCGG	Bsp120I
	(SEQ ID NO:103)		NotI

- 239 -

JC12	(SEQ ID NO:104)	3' GT ACT TCG CATG	Acc56I BsrGI
JC14	(SEQ ID NO:105)	3' GT ACT TCG AGCT	SalI .
JC15	(SEQ ID NO:106)	3' GT ACT TCG ACGT	Ppu10I ApaLI

Tables 3 and 4 list the RE combinations that have been tested in QEA™ method experiments on human placental and glandular cDNAs samples. The preferred double digests are those that give more than approximately 50 bands in the range of 100 to 700 bp. Table 11 lists the preferred RE combinations for human cDNA analyses.

TABLE 3:   PREFERRED RE COMBINATIONS FOR HUMAN cDNA ANALYSIS

Acc56I & HindIII	Acc65I & NgoMI	BamHI & EcoRI
BglII & HindIII	BglII & NgoMI	BsiWI & BspHI
BspHI & BstYI	BspHI & NgoMI	BsrGI & EcoRI
EagI & EcoRI	EagI & HindIII	EagI & NcoI
HindIII & NgoMI	NgoMI & NheI	NgoMI & SpeI
BglII & BspHI	Bsp120I & NcoI	BssHII & NgoMI
EcoRI & HindIII	NgoMI & XbaI	

Table 4 lists other RE combinations tested and that can be used for human cDNA analyses.

TABLE 4:   OTHER RE COMBINATIONS FOR HUMAN cDNA ANALYSIS

AvrII & NgoMI	BamHI & Bsp120I	BamHI & BspHI
BamHI & NcoI	BclI & BspHI	BclI & NcoI
BglII & BspEI	BglII & EcoRI	BglII & NcoI
BssHII & BsrGI	BstYI & NcoI	BamHI & HindIII
BglII & Bsp120I	BspHI & HindIII	

Tables 5 and 6 list the RE combinations that have been tested in QEA™ method experiments on mouse cDNA samples.

- 240 -

The preferred double digests are those that give more than
approximately 50 bands in the range of 100 to 700 bp.   Table
5 lists the preferred RE combinations for mouse cDNA
analyses.

5

TABLE 5:   PREFERRED RE COMBINATIONS
FOR MOUSE cDNA ANALYSIS

Acc56I & HindIII	Acc65I & NgoMI	AscI & HindIII
AvrII & NgoMI	BamHI & BspHI	BamHI & HindIII
BamHI & NcoI	BclI & NcoI	BglII & BspHI
BglII & HindIII	BglII & NcoI	BglII & NgoMI
Bsp120I & NcoI	Acc65I & BspHI	BspHI & Bsp120I
BspHI & BsrGI	BspHI & EagI	BspHI & NgoMI
BspHI & NotI	BssHII & HindIII	BstYI & HindIII
HindIII & NcoI	HindIII & NgoMI	NcoI & NotI
NgoMI & NheI	NgoMI & SpeI	NgoMI & XbaI
BclI & HindIII		

20
Table 6 lists other RE combinations tested and that
can be used for mouse cDNA analyses.

TABLE 6:   OTHER RE COMBINATIONS FOR MOUSE cDNA ANALYSIS

Acc65I & NcoI	BclI & BspHI	BsiWI & BspHI
BsiWI & NcoI	BspHI & HindIII	BsrGI & NcoI
BssHII & NgoMI	BstYI & BspHII	EagI & NcoI
HindIII & MluI		

30
Table 7 lists the data obtained from various RE
combinations using mouse cDNA samples.   The number of bands
was observed from silver stained acrylamide separation gels.

TABLE 7:   MOUSE cDNA RE DIGESTION RESULTS

35

RE Combination	Number of Bands

Acc65I & HindIII	200
Acc65I & NgoMI	150
AscI & HindIII	100
AvrII & NgoMI	50
BamHI & BspHI	200
BamHI & HindIII	150
BamHI & NcoI	150
BclI & BspHI	5
BclI & HindIII	150
BclI & NcoI	50
BglII & BspHI	50
BglII & HindIII	150
BglII & NcoI	50
BglII & NgoMI	50
Bsp120I & NcoI	50
BspHI & Acc65I	150
BspHI & Bsp120I	50
BspHI & BsrGI	200
BspHI & EagI	150
BspHI & HindIII	0
BspHI & NgoMI	150
BspHI & NotI	150
BsrGI & NcoI	10
BssHII & HindIII	100
BssHII & NgoMI	20
BstYI & BspHI	20
BstYI & HindIII	200
EagI & NcoI	10
HindIII & MluI	25
HindIII & NcoI	50
HindIII & NgoMI	150
NcoI & NotI	200

NgoMI & NheI	50
NgoMI & SpeI	200
NgoMI & XbaI	50
TOTAL # BANDS	3490

31 available REs that recognize a 6 bp recognition sequence and generate a 4 bp 5' overhang are: Acc65I, AflII, AgeI, ApaLI, ApoI, AscI, AvrI, BamHI, BclI, BglII, BsiWI, Bsp120I, BspEI, BspHI, BsrGI, BssHII, BstYI, EagI, EcoRI, HindIII, MfeI, MluI, NcoI, NgoMI, NheI, NotI, Ppu10I, SalI, SpeI, XbaI, and XhoI.

All of these enzymes have been tested in QEA™ method protocols with the specified buffer conditions with the exception of AflII. All were useable except for MfeI, Ppu10I, SalI, and XhoI. All the other 26 enzymes have been tested and are usable in the RE implementation of QEA™ method.

However certain pairs of these enzymes are less informative due to the fact that they produce identical overhangs, and thus their recognition sequences cannot be distinguished by the QEA™ method adapters. These pairs are Acc65I and (BsiWI or BsrGI); AgeI and (BspEI or NcoMI); ApoI and EcoRI; AscI and (BssHII or MluI); AvrI and (NheI, SpeI, or XbaI); BamHI and (BclI, BglII, or BstYI); BclI and (BgLII or BstYI); BglII and BstYI; BsiWI and BsrGI; Bsp120I and EagI; BspEI and NcoMI; BspHI and NcoI; BssHII and MluI; NheI and (SpeI or XbaI); and SpeI and XbaI.

Thus, 301 RE pairs have been tested and are useable in the RE embodiments of the QEA™ method.

6.1.12.4. FLUORESCENT LABELS FOR QEA™ METHODS

Fluorochromes labels that can be used in QEA™ methods include the classic fluorochromes as well as more specialized fluorochromes. The classic fluorochromes include bimane, ethidium, europium (III) citrate, fluorescein, La Jolla blue, methylcoumarin, nitrobenzofuran, pyrene butyrate,

rhodamine, terbium chelate, and tetramethylrhodamine. More
specialized fluorochromes are listed in Table 8 along with
their suppliers.

TABLE 8: FLUORESCENT LABELS

Fluorochrome	Vendor	Absorption Maximum	Emission Maximum
Bodipy 493/503	Molecular Probes	493	503
Cy2	BDS	489	505
Bodipy FL	Molecular Probes	508	516
FTC	Molecular Probes	494	518
FluorX	BDS	494	520
FAM	Perkin-Elmer	495	535
Carboxyrhodamine	Molecular Probes	519	543
EITC	Molecular Probes	522	543
Bodipy 530/550	Molecular Probes	530	550
JOE	Perkin-Elmer	525	557
HEX	Perkin-Elmer	529	560
Bodipy 542/563	Molecular Probes	542	563
Cy3	BDS	552	565
TRITC	Molecular Probes	547	572
LRB	Molecular Probes	556	576
Bodipy LMR	Molecular Probes	545	577
Tamra	Perkin-Elmer	552	580
Bodipy 576/589	Molecular Probes	576	589
Bodipy 581/591	Molecular Probes	581	591
Cy3.5	BDS	581	596
XRITC	Molecular Probes	570	596
ROX	Perkin-Elmer	550	610
Texas Red	Molecular Probes	589	615
Bodipy TR (618?)	Molecular Probes	596	625
Cy5	BDS	650	667
Cy5.5	BDS	678	703

Fluorochrome	Vendor	Absorption Maximum	Emission Maximum
DdCy5	Beckman	680	710
Cy7	BDS	443	767
DbCy7	Beckman	790	820

The suppliers listed in Table 8 are Molecular Probes (Eugene, OR), Biological Detection Systems ("BDS") (Pittsburgh, PA) and Perkin-Elmer (Norwalk, CT).

Means of utilizing these fluorochromes by attaching them to particular nucleotide groups are described in Kricka et al., 1995, Molecular Probing, Blotting, and Sequencing, chap. 1, Academic Press, New York. Preferred methods of attachment are by an amino linker or phosophoramidite chemistry.

6.1.12.5.   PREFERRED REACTANTS FOR SEQ-QEA™ METHODS

Table 9 lists exemplary Type IIS REs adaptable to a SEQ-QEA™ method and their important characteristics. For each RE, the table lists the recognition sequence on each strand of a dsDNA molecule and the distance in bp from the recognition sequence to the location of strand cutting. Also listed is the net overhang generated.

TABLE 9: SAMPLE TYPE IIS REs

RE	Recog. Seqs.	Dist. to cutting site (bp)	Over-hang (bp)	Comment
FokI	GGATG CCTAC	9 13	4	
HgaI	GACGC CTGCG	5 10	5	
BbvI	GCAGC CGTCG	8 12	4	
BsmFI	GGGAC CCCTG	10 14	4	Lower recognition site specificity

BspMI	ACCTGC TGGACG	4 8	4	
SfaNI	GCATC CGTAG	5 9	4	

Table 10 lists exemplary primer and linker combinations adaptable to a SEQ-QEA™ method. They satisfy the previously described requirements on primers and linkers. Except for the indicated differences, they are the same as the primers and linkers of similar names in Table 10. RA24-U and RC24-U have a 5′ biotin capture moiety and a uracil release means as indicated, and are adaptable to the same linkers and REs as are RA24 and RC24 of Table 10. RA24-S and RC24-S also have a 5′ biotin capture moiety with a AscI recognition site release means as indicated, and are adaptable to the same linkers and REs as are RA24 and RC24 of Table 10. JA24-K has an internal FokI recognition site as indicated and a 5′ FAM label moiety (see Table 16). Linkers KA5, KA6, and KA9 with the indicated REs function with this primer. JC24-B has an internal BbvI recognition site, a 5′ FAM label, and functions with linkers BA5 and BA9.

TABLE 10:   SAMPLE ADAPTERS

Series	Adapter: Primer (longer strand) Linker (shorter strand) Notes: 'b' signifies a biotin moiety 'f' signifies a FAM label moiety	RE
RA24-U	5′ b-AGC ACT CTC CAG CC**U** CTC ACC GAA 3′ (SEQ ID NO:107)	
RA24-S	5′ b-AGC ACT CT**G GCG CGC C**TC ACC GAA 3′ (SEQ ID NO:108)	
RC24-U	5′ b-AGC ACT CTC CAG CC**U** CTC ACC GAC 3′ (SEQ ID NO:109)	
RC24-S	5′ b-AGC ACT CT**G GCG CGC C**TC ACC GAC 3′ (SEQ ID NO:110)	

JA24-K	5′ f-ACC GAC GTC GAC TAT **GGA TG**A AGA 3′ (SEQ ID NO:111)	FokI (9)
KA9	3′ CT ACT TCT CTAG (SEQ ID NO:112)	DpnII BglII BamHI BclI BstYI
KA5	3′ CT ACT TCT GTAC (SEQ ID NO:113)	NcoI BspHI
KA6	3′ CT ACT TCT GGCC (SEQ ID NO:114)	AgeI NgoMI BspEI SgrAI BsrFI BsaWI
JC24-B	5′ f-ACC GAC GTC GAC TAT C**GC AGC** 3′ (SEQ ID NO:115)	BbvI (8)
BA9	3′ CG TCG TCT CTAG (SEQ ID NO:116)	DpnII BglII BamHI BclI BstYI
BA5	3′ CG TCG TCT GTAC (SEQ ID NO:117)	NcoI BspHI

6.1.13.   POST-MATING VERIFICATION PROTOCOLS

It is advantageous to perform verification protocols on yeast colonies that have been selected as positive for protein-protein interactions. Such protocols can further screen out both falsely positive colonies as well as eliminate non-specific protein-protein interactions. A non-specifically interacting protein fragment is one that interacts indiscriminately with many other protein fragments, and thereby, is unlikely to be biologically significant. The remaining yeast colonies should represent true and specific protein-protein interactions.

6.1.13.1.   PLASMID DROP-OUT PROTOCOL

The plasmid drop-out protocol, performed after selection for protein-protein interaction, further screens

out colonies that are falsely positive for protein-protein
interaction due to fortuitous activation of reporter genes by
a non-interacting binding domain fusion protein. Pre-mating
negative selection, even according to the rigorous protocol
5 of Section 6.1.7, does not screen out all fortuitously
activating binding domain fusions. The more complex the
binding domain library, the more such fortuitously activating
fusions escape such initial selection. For binding domain
fusion libraries with a complexity of 10^6, or 10^7, or greater,
10 post-mating screening is especially preferred.

In summary, the plasmid drop-out protocol applied
to a colony positive for protein-protein interaction, first,
selects for progeny that have lost either one of the
activation domain or binding domain plasmids, and second,
15 checks these progeny yeast, bearing only one of the plasmids,
for activation of reporter gene(s). If a reporter gene is
activated in a yeast progeny bearing only a single plasmid,
the original colony is falsely positive for interaction. In
all cases, false positives due to fortuitous activation by
20 binding domain fusions are preferably checked. False
positives due to fortuitous activation by activation domain
fusions are not routinely checked since such fortuitous
activation has only been very rarely observed. Accordingly,
this protocol is described to check for fortuitous activation
25 by binding domain fusions. Adaption of the steps to check
activation domain fusions will be apparent to one of skill in
the art.

In a specific example, the plasmids with binding
domain fusions express TRP1, the plasmids with activation
30 domain fusions express LEU2, and *lacZ* is a reported gene.
Adaption of the steps to check other combinations of
selectable markers will be apparent to one of skill in the
art.

In detail, yeast cells are selected for plasmid
35 drop-out by growth on a rich, non-selective medium. Yeast
cells from colonies positive for interaction are inoculated
into 2 ml of a rich medium like YPAD in 15 ml test tubes and

allowed to grow with gentle agitation at 30° C for two days or until stationary phase. A 30 μl aliquot of a 10⁻⁴ dilution of this culture is plated on a first 100 mm plate that has medium selective only for the binding domain plasmid, and
5 allowed to grow for 1-2 days. Second, plasmid drop-out is assayed by replica plating colonies from this first plate onto two selection plates, one with a medium selective for the binding domain plasmid and the other with medium selective for the activation domain plasmid. The yeast cells
10 on the two selection plates are allowed to grow for an additional 1-2 days or until colony growth becomes visible. The plate selective for the binding domain fusion is also assayed for reporter gene activity by, *e.g.*, the filter-lift assay of Section 6.1.11 for β-galactosidase activity, where
15 *lacZ* is one of the reporter genes. Colonies which grow on the plate selective for the binding domain plasmid but not on the plate selective for the activation domain plasmid have dropped the activation domain plasmid. Any of these latter colonies which are also positive for reporter gene activity
20 are false positive. In these colonies the binding domain fusion protein alone has fortuitously activated the reporter gene(s). These false-positives are discarded from further consideration.

25   6.1.13.2. <u>YEAST MATRIX-MATING PROTOCOL</u>
          The yeast matrix-mating protocol, also performed after colonies have been selected for protein-protein interaction, eliminates positive colonies due to non-specific protein-protein interactions. Although a colony observed to
30 be positive for protein-protein inverations may have reporter gene activation due to true protein-protein association, this association may be non-specific and not of particular interest (the protein participating in such non-specific interactions being referred to herein as a "sticky" protein).
35 For example, either the binding or activation domain fusion protein may bear a fragment capable of associating with a wide range of, *e.g.*, hydrophobic domains on many other

- 249 -

activation or binding domain, respectively, fusion proteins
and thereby activate reporter genes. Such non-specific
association may be of less interest than specific binding
between proteins that may represent, e.g., a portion of an
5 intracellular signaling pathway. The matrix-mating protocol
finds those activation or binding domain fusion proteins that
associate non-specifically with many other partners in a
particular mating.

In summary, the matrix mating protocol performs a
10 second, limited two-hybrid mating using only activation and
binding domain plasmids from colonies that were true
positives in the first and original two-hybrid mating. For
example, if the first, original two-hybrid screen has an M x
N (M and N representing the complexity in the two different
15 populations used to make fusion contructs) complexity of say
10^7 x 10^7 and found, e.g., 50 to 100 interacting pairs, then
the second screen can have a complexity of from 50 x 50 to
100 x 100. If a particular fusion protein participates only
in specific protein-protein interactions, then in the second
20 mating, it is likely that the only positive mating will be
that with the same other fusion protein that was positive in
the original mating. On the other hand, if the particular
fusion protein binds non-specifically, then in the second
mating, it is likely that matings with many, perhaps all, of
25 the other fusion proteins will be positive. Preferably,
matrix mating is performed only on colonies positive for
interaction that have passed the plasmid drop-out test.

The matrix-mating protocol is adapted to the
limited nature of the second mating. First, DNA is extracted
30 from colonies found to be positive for protein-protein
interaction in the first mating; second, yeast strains of
opposite mating type are transformed with the binding and
activation domain plasmids rescued from the extracted DNA;
and third, the transformed yeast strains are mated and
35 screened for protein-protein interaction. Alternatively, the
matrix-mating two-hybrid screen can be performed according to
the protocols of the first mating, as previously described.

The first step, preferably, extracts DNA by binding
it to magnetic beads or a similar substrate, such as BioMag
beads, catalog No. 8-MB4125B from PerSeptive BioSystems
(Boston, MA). An aliquot of 150 μl of cells from a colony
5 positive for interaction is pelleted for 3 minutes at 3500
rpm. The pellet is resuspended in 40 μl of Z-buffer
containing 300 μg/ml of Zymolase, and incubated at 37° C for
1 hour. (Z-buffer is made by adding to 800 ml of water 16.1
g of Na_2HPO_4, 5.5 g of NaH_2PO_4, 0.75 g of KCl, and 0.246 g of
10 $MgSO_4.7H_2O$, adjusting the pH to 7.0, and adding water to 1000
ml.) The cell debris are spun down, and the supernatant
transferred to a new tube. An 40 μl aliquot of binding
buffer (2.5 M MgCl in 20% PEG having a molecular weight of
approximately 8000) and 10 μl of pre-washed BioMag beads are
15 added to the supernatant and incubated at room temperature
for 5-10 minutes. Finally, the beads are precipitated with a
magnetic bed and washed twice with washing buffer (70% EtOH,
30% 10 mM Tris with 1 mM EDTA). DNA is eluted from the
washed beads in 10 μl of TE buffer.
20   In the second step, plasmids in the extracted DNA
are rescued into *E. coli* according to protocols known in the
art, such as that found in Sambrook et al., 1989, Molecular
Cloning, A Laboratory Manual, Cold Spring Harbor Press, Cold
Spring, N.Y., which is incorporated here in its entirety by
25 reference. *E. coli* bearing the rescued plasmids are
maintained in media selective for the particular plasmid, as
by containing an antibiotic whose resistance is coded for by
a gene on the plasmid expressed in *E. coli*. Yeast strains
are transformed with the plasmids rescued into the *E. coli*
30 according to protocols known in the art, such as that found
in Sambrook et al., *supra*. All the activation and binding
domain plasmids are transformed into yeast strains of
opposite mating type. The yeast strains transformed with the
plasmids are maintained in media appropriately selective for
35 the particular plasmid, as by lacking a particular nutrient
whose synthesis is coded for by a gene on the plasmid
expressed in yeast.

Finally, in the third step, the transformed are mated. Cells from each of the yeast strains individually transformed, by way of example, with binding domain plasmids from one of the originally positive colonies are suspended in

5 separate 50 µl aliquots of sterile water. Aliquots of 20 µl of the resuspended cells are uniformly seeded along a separate straight lines on plates appropriately selective for the binding domain plasmid, and the plate is incubated overnight at 30° C. Next, cells from each of the yeast

10 strains individually transformed with activation domain plasmids from one of the originally positive colonies are similarly seeded along straight lines on a plate with a rich medium like YPAD. Mating is performed by replica plating the plate bearing the binding domain transformants onto the YPAD

15 plate bearing the activation domain transformants in such a manner that the two sets of seeding lines are approximately at right angles to each other, and by overnight incubation at 30° C. Finally, colonies having protein-protein interactions are assayed for by replica plating the YPAD mating plate onto

20 an assay plate selective both for the activation and binding domain plasmids and for the reporter genes activated by protein-protein interaction. Plasmid drop-out can also be checked for by replica plating onto a plate selective only for the two plasmids.

25   The assay plate indicates specificity of protein-protein interactions. A specifically interacting protein is represented by growth on the assay plate only at the intersection of its seeding line with the seeding line of yeast transformed with its interacting partner observed in

30 the original mating. The intersection of these two seeding lines reconstitutes the originally observed interaction. A non-specifically interaction protein is represented by growth at many, perhaps all, of the intersections of its seeding line of with the seeding lines of the other yeast

35 transformants. Thereby, matrix mating distinguishes specific and non-specific protein-protein interactions in the colonies positive for interaction in the original mating.

6.2.   LIBRARIES

Libraries in pAD-GAL and in pBD-GAL or pAS2-1 of ~ 1 x 10⁶ clones were made from 1-10 μg of cDNA from a colon cancer cell line by the methods described above in Section

5 6.1.6.   The libraries were propagated in the *E. coli* strain XL1Blue (Stratagene) and plasmid DNA was extracted by standard procedures.

6.3. CONSTRUCTION OF YEAST STRAINS

10 **Construction of reporter systems**

The Reporter System is binary in nature and consists of two halves with each half containing a reporter strain.   Each half is of the opposite mating type, *i.e.*, a or α.   In a preferred embodiment, the mating type a reporter

15 strain contains an intrachromosomal *URA3* Reporter Gene under the control of the *GAL1* promoter and its native GAL4 DNA binding sites; and the mating type α reporter strain contains both an intrachromosomal *HIS3* Reporter Gene and an intrachromosomal *lacZ* Reporter Gene, each under the control

20 of the *GAL1* promoter and its native GAL4 DNA binding sites.

The a strain YULH contains the *URA3* Reporter Gene under the control of a promoter that contains GAL4 binding sites.

The α strain N106' contains two reporters: a *HIS3*

25 Reporter Gene under the control of a *HIS3* promoter that has been engineered to contain GAL4 binding sites, and a *lacZ* Reporter Gene under the control of a *GAL1* promoter.

The a strain N105' contains two reporters: a *HIS3* Reporter Gene under the control of a *HIS3* promoter that has

30 been engineered to contain GAL4 binding sites, and a *lacZ* Reporter Gene under the control of a *GAL1* promoter.

The a strain N105 contains two reporters: a *HIS3* Reporter Gene under the control of a *HIS3* promoter that has been engineered to contain GAL4 binding sites, and a *lacZ*

35 Reporter Gene under the control of a *GAL1* promoter.   The strain is not deficient in *LYS2* or *URA3*.

- 253 -

The α strain N106 contains two reporters: a *HIS3*
Reporter Gene under the control of a *HIS3* promoter that has
been engineered to contain GAL4 binding sites, and a *lacZ*
Reporter Gene under the control of a *GAL1* promoter. The
5 strain is not deficient in *LYS2* or *URA3*.

In one embodiment of the invention, the two
reporter strains are N105' (mating type a) and N106' (mating
type α).

In another embodiment of the invention, the two
10 reporter strains are YULH (mating type a) and N106' (mating
type α). In another embodiment, N203 is used as mating type
a. N105 (which is not *ura3* or *lys2*) can be used as an
alternative to YULH if uracil selection is not desired for
use. Details of the methods used to construct these strains
15 are presented in the subsections below.

6.3.1. <u>CONSTRUCTION OF STRAINS N105 AND N106</u>

Strains N105 and N106 were derived from the strain
Y190 (available from Clontech; Harper et al., 1993, Cell
20 75:805-816). The a strain Y190 contains two reporters: a
HIS3 Reporter Gene under the control of a *HIS3* promoter that
has been engineered to contain GAL4 binding sites, and a *lacZ*
Reporter Gene under the control of a *GAL1* promoter. Y190 (a
gift of Stephen J. Elledge, Baylor College of Medicine) was
25 diploidized by transforming it with a plasmid bearing a copy
of the HO gene (Herskowitz and Jensen, 1991, Meth. Enzymol.
194:132-146). The HO gene switches the mating type of the
strain and thus, when two opposite mating types exist, they
mate to form diploids. The diploids were then transferred to
30 sporulating medium on plates (Sherman et al., eds., 1991,
Getting started with yeast, Vol. 194, Academic Press, New
York) and left to sporulate at 30°C for 2 days. The haploids
were isolated by dissection of tetrads and the two mating
types were determined by mating to tester a and α strains, a
35 will not mate with a, and α will not mate with α. These two
strains, with the exception of being opposite mating types,

are truly isogenic and the genotype includes *leu2, trp1, his3, URA3::GAL-lacZ, LYS2::GAL-HIS3.*

6.3.2. CONSTRUCTION OF THE REPORTER STRAIN N106'

The strain N106 was made deficient in *ura3* by selection of ura minus cells on 5-FOA plates. Then, these cells were made *lys2* (lysine minus) by a two-step gene-disruption method (Rothstein, 1983, Methods. Enzymol. 101:202-211), so that, if desired, a *LYS2* Reporter Gene or a plasmid containing *LYS2* can be selected for in the strain. A mutant version of the *lys2* gene was used for this purpose. This mutant *lys2-ΔNheI* (a gift of Albert Smith, Yale University) was generated by deleting the NheI fragment that is internal to the *LYS2* coding region (Fleig et al., 1986, Gene 46:237-245). This gene is in a plasmid that was linearized with XbaI and the linearized DNA was used to transform N102 by the lithium acetate transformation protocol of Section 6.1.2. This plasmid is also marked with *URA3* and so cells in which the plasmid had integrated were selected on ura minus plates. These transformants were then patched out onto 5-FOA plates and ura minus cells were recovered. These ura minus cells were patched out simultaneously onto lysine minus plates and YPAD plates, and cells that did not grow on the lysine minus plates were chosen. In this manner, cells that were *lys2* were recovered and the strain was named N106'. The genotype of this strain is *MATα, ura3, his3, lys2, ade2, trp1, leu2, gal4, gal80, cyh^r, lys2::GAL1$_{UAS}$-HIS3$_{TATA}$- HIS3, ura3::GAL1$_{UAS}$-GAL$_{TATA}$-lacZ.*

6.3.3. CONSTRUCTION OF THE REPORTER STRAIN N105'

The strain N105 was made deficient in *ura3* by selection of ura minus cells on 5-FOA plates. Then, these cells were made *lys2* (lysine minus) by a two-step gene-disruption method (Rothstein, 1983, Methods. Enzymol. 101:202-211), so that, if desired, a *LYS2* Reporter Gene or a plasmid containing *LYS2* can be selected for in the strain.

A mutant version of the *LYS2* gene was used for this purpose. This mutant *lys2-ΔNheI* (a gift of Albert Smith, Yale University) was generated by deleting the NheI fragment that is internal to the *LYS2* coding region (Fleig et al., 1986,

5 Gene 46:237-245). This gene is in a plasmid that was linearized with XbaI and the linearized DNA was used to transform N101 by the lithium acetate transformation protocol of Section 6.1.2. This plasmid is also marked with *URA3* and so cells in which the plasmid had integrated were selected on

10 ura minus plates. These transformants were then patched out onto 5-FOA plates and ura minus cells were recovered. These ura minus cells were patched out simultaneously onto lysine minus plates and YPAD plates, and cells that did not grow on the lysine minus plates were chosen. In this manner, cells

15 that were *lys2* were recovered and the strain was named N105'. The genotype of this strain is *MATa, ura3, his3, lys2, ade2, trp1, leu-2, gal4, gal80, cyh^r, lys2::GAL1$_{UAS}$-HIS3$_{TATA}$-HIS3, ura3::GAL1$_{UAS}$-GAL$_{TATA}$-lacZ.*

20   6.3.4.   <u>CONSTRUCTION OF THE REPORTER STRAIN YULH</u>
                The strain Y166 (a gift of Stephen J. Elledge, Baylor College of Medicine) was made *lys2* (lysine minus) by a two-step gene-disruption method (Rothstein, 1983, Methods. Enzymol. 101:202-211), so that, if desired, a *LYS2* Reporter

25 Gene or a plasmid containing *LYS2* can be selected for in the strain. A mutant version of the *LYS2* gene was used for this purpose. This mutant *lys2-ΔNheI* (a gift of Albert Smith, Yale University) was generated by deleting the NheI fragment that is internal to the *LYS2* coding region (Fleig et al.,

30 1986, Gene 46:237-245). This gene is in a plasmid that was linearized with XbaI and the linearized DNA was used to transform N101 by the lithium acetate transformation protocol of Section 6.1.2. This plasmid is also marked with *URA3* and so cells in which the plasmid had integrated were selected on

35 ura minus plates. These transformants were then patched out onto 5-FOA plates and ura minus cells were recovered. These ura minus cells were patched out simultaneously onto lysine

minus plates and YPAD plates, and cells that did not grow on
the lysine minus plates were chosen.   In a similar manner,
these cells that were *lys2* were also made *his3* (histidine
minus) by a two-step gene disruption method.   A mutant

5 *his3-NdeI* (a gift of Petra RossMacDonald, Yale University)
was used for this purpose.   This mutant *his3-Nde I* was
generated by digesting the *HIS3* gene in the plasmid pRS303
(Sikorski and Heiter, 1989, Genetics 122:19-27) and filling
in the NdeI site with Klenow DNA Polymerase I and dNTPs.

10 Then the *URA3* gene was removed as a EagI-SmaI fragment from
the plasmid YiP5 (Struhl et al., 1979, Proc. Natl. Acad. Sci.
72:1035-1039) and cloned in between the same sites in pRS303.
This plasmid was linearized with NheI and the linearized DNA
was used to transform the Y166 derivative that is *lys2*, by

15 the lithium acetate transformation protocol of Section 6.1.2.
This plasmid is also marked with *URA3* and so cells in which
the plasmid had integrated were selected on ura minus plates.
These transformants were then patched out onto 5-FOA plates
and ura minus cells were recovered.   These ura minus cells

20 were patched out simultaneously onto histidine minus plates
and YPAD plates, and cells that did not grow on the histidine
minus plates were chosen.   In this manner, cells that were
his3 were recovered and the strain was named YULH.   The
genotype of this strain is *MATa, ura3, his3, lys2, ade2,*

25 *trp1, leu2, gal4, gal80, GAL1-URA3*.

6.3.5.   CONSTRUCTION OF THE YEAST STRAIN N203

This section describes methods for the construction
of a yeast strain, termed N203, bearing a *URA3* Reporter Gene

30 under the control of a *GAL1-10* promoter (driven by GAL4 DNA
binding sites), that can be used in place of strain YULH for
detecting protein-protein interactions.

Construction of the *GAL1-10::URA3* fusion gene

35   The *GAL1-10* promoter (Yocum et al. 1984, Mol. Cell.
Biol. 4:1985-1998) is used to create the *GAL1-10::URA3*

fusion gene.   The *GAL1-10* promoter is isolated by PCR from
yeast genomic DNA by using the following oligonucleotides:
G1

5'-GAGAGAGAGAG<u>GGTACC</u>GAACCAATGTATCCAGCACCACCTGTAACC-3'
                  KpnI
5 (SEQ ID NO:39)

G2

5'-GAGAGAGAG<u>GAATTC</u>CATTATAGTTTTTTCTCCTTGACGTTAAAGTATAGAGG-3'
        EcoRI

10 (SEQ ID NO:40)

     The two primers flank the entire *GAL1-10* promoter
(Yocum et al., 1984, Mol. Cell. Biol. 4:1985-1998).   The two
primers also donate the restriction sites KpnI and EcoRI.
The *GAL1-10*-specific sequences are italicized.   The primer G1
15 contains the sequences of the *GAL10* coding region from
position +74 to +44, with +1 being the start of the coding
region.   The primer G2 contains the ATG codon of the *GAL1*
gene and the 35 nucleotides upstream to it.   The PCR products
are digested with KpnI and EcoRI and cloned between the same
20 sites in the plasmid SK+ (Stratagene) to yield the plasmid
GAL1-SK.

     The *URA3* gene is amplified by PCR using the
following oligonucleotides and yeast genomic DNA as template:
5'-GAGAGAGA<u>GAATTCT</u>CGAAAGCTACATATAAGGAACGTGCTGC-3' (SEQ ID NO:41)
25   EcoRI

5'-GAGAGAGA<u>CGGCCG</u>CGTCATTATAGAAATCATTACGACCGAG-3' (SEQ ID NO:42)
        EagI

     The *URA3*-specific sequences are italicized and the
30 *URA3* sequences extend from the second codon to the 3' end of
the gene. The PCR products are digested with EcoRI and EagI
and cloned between the same sites in GAL1-SK. This creates a
GAL1-10::URA3 fusion that contains all of the *URA3* protein
except the first ATG and also contains the ATG of *GAL1*.   Two
35 amino acids (glutamate and phenylalanine) are added at the
junction of *GAL1* and *URA3* by the cloning protocol (*i.e.*, by
the addition of the EcoRI recognition site).   The

GAL1-10::URA3 fusion has GAL$_{UAS}$ in its promoter and so can be activated by the GAL4 protein.

Construction of the Yeast Strain
Bearing the *GAL1-10::URA3* Reporter Gene

The *GAL1-10::URA3* is amplified using the following oligonucleotides:

U1

5'-GATTCGGTAATCTCCGAACAGAAGGAAGAACGAAGGAAGGAGCACAGACTTAGATTGG
TAGAACCAATGTATCCAGCACCACCTGTAACC-3' (SEQ ID NO:43)

U2

5'-ACATCAAAAGGCCTCTAGGTTCCTTTGTTACTTCTTCCG-3' (SEQ ID NO:44)

The oligonucleotide U1 contains the 60 nucleotides (+67 to +126) of the *URA3* sequence upstream of the promoter (Rose et al., 1984, Gene 29:113-124) fused to the 30 nucleotides of the *GAL1-10* promoter (italicized; Yocum et al. 1984, Mol. Cell. Biol. 4:1985-1998). The oligonucleotide U2 contains sequences from within the coding region (+632 to 670; Rose et al., 1984, Gene 29:113-124). *GAL1-10::URA3* is used as the template for the PCR reaction.

The strain N201 contains copies of the RAS-GBD and RAF-GAD plasmids described in Section 6.4 and is derived by the transformation of the strain N200 that is itself a derivative of the strain CG-1945 (Clontech Laboratories, Inc., Palo Alto, CA) with the RAS-GBD and RAF-GAD plasmids. The genotype of the CG-1945 strain is *MATa, ura3-52, his3-200, lys2-801, ade2-101, trp1-901, leu2-3, 112, gal4-542, gal80-538, cyh^{r}2, LYS2::GAL1$_{UAS}$-GAL1$_{TATA}$-HIS3, URA3::GAL1$_{UAS\ 17mers\ x(3)}$-CYC1$_{TATA}$-lacZ*. N200 is obtained by selecting ura minus cells by 5-FOA resistance selection. This is performed by patching cells onto 5-FOA plates. The RAS-GBD and RAF-GAD transformants of N200 are selected on SC-TRP and SC-LEU plates respectively, as the RAS-GBD and the RAF-GAD plasmids are marked with *TRP1* and *LEU2* genes, respectively.

The PCR products from a reaction using the oligonucleotides U1 and U2 are used to transform the yeast strain N201, and the transformants are selected on SC-TRP-LEU-URA medium. The GAL1-10::URA3 gene is inserted at

5 the ura3-52 locus by a double reciprocal recombination event (Rothstein, 1983, Methods. Enzymol. 101:202-211). The interaction of the RAS-GBD and RAF-GAD plasmids reconstitutes the transcriptional activator GAL4 that activates transcription from the GAL1-10::URA3 gene, thereby expressing

10 the GAL1::URA3 fusion reporter gene and thus, enabling the cells to grow in the absence of uracil. The dependence of the +URA phenotype on the reconstitution of GAL4 is confirmed by the inability of cells, that have lost the RAS-GBD and RAF-GAD plasmids, to grow in the absence of uracil. This

15 derivative of N201 bearing the GAL1-10::URA3 gene and the RAS-GBD and RAF-GAD plasmids is named N202.

The strain N202 is streaked out on YPAD plates and individual colonies that have lost both the RAS-GBD and the RAF-GAD plasmids are selected by their inability to grow on

20 media lacking either tryptophan or leucine, respectively. This strain is named N203 and is a strain bearing the GAL1-10::URA3 Reporter Gene that can be used for detecting protein-protein interactions.

The strain N203 can be transformed with both the

25 GBD and GAD plasmids to detect protein-protein interactions. Alternatively, this strain bearing just one of the plasmids (GBD or GAD) can be mated to another strain like N106' that bears the other kind of plasmid (GBD or GAD). Since the N203 strain has the URA3 Reporter Gene, it can be used for

30 counterselection on 5-FOA plates to eliminate the false positives that may arise from the activation of the URA3 reporter gene by the GBD plasmid alone.

35

**Counter Selection of N203 Transformants
on 5-FOA Plates to Eliminate False-Positives**

The strain N203 is transformed with the pAS2-1
library and selected with 5-FOA as described in Section
6.1.7.

6.4. CONSTRUCTION OF FUSION GENES

The pairs of interacting proteins, against which
peptide inhibitors are to be screened, were introduced into
the reporter strains as fusion genes. RAS was introduced as
a GAL4 DNA-Binding Domain fusion (GBD), termed RAS-GBD, RAF
was introduced as a GAL4 Activation Domain fusion (GAD)
termed RAF-GAD, Vascular Endothelial Growth Factor (VEGF) was
introduced both as a GAL4 DNA-Binding domain fusion protein
(VEGF-GBD) and GAL4 activation domain fusion protein (VEGF-
GAD), and KDR (receptor for VEGF) was introduced as a GAL4
activation domain fusion protein (KDR-GAD). The complete RAS
protein was used in making the fusion (Miura et al., 1986,
Jpn. J. Cancer Res. 77:45-51), the RAF sequences extend from
amino acids 1 to 257 of the RAF protein (Bonner et al., 1986,
Nucleic Acids Res. 14:1009-1015), the VEGF sequences extend
from amino acids 32 to the C terminus of the protein of the
VEGF-165 protein (Leung et al., 1989, Science 246:1306-1309),
and the KDR sequences extend from amino acids 19 to 757 of
the KDR protein (Terman et al., 1992, Biochem. Biophys. Res.
Comm. 187:1579-1586).

The plasmid vectors for the GBD fusions and the GAD
fusions, pAS2 and pACT2, respectively (Clontech) were each
modified to introduce two SfiI sites to facilitate cloning of
insert DNAs. These plasmids are yeast *E. coli* shuttle
vectors and are marked with β-lactamase for selection in *E.
coli* using ampicillin and a 2μ circle DNA for replication in
yeast. The pAS2 plasmid (Clontech; also known as pAS1-CYH,
Harper et al., 1993, Cell 75:805-816) is marked with the *TRP1*
gene for selection in yeast (in medium lacking tryptophan)
whereas the pACT2 is marked with the *LEU2* gene for the same
(in medium lacking leucine). The resulting plasmids with the

two SfiI sites were called pASSfiI and pACTSfiI,
respectively.   The polylinkers of the plasmids are as
follows:

pASSfiI:

5
```
            SfiI                    SfiI
5'-CAT ATG GCC GAG GTG GCC TAG GGC CTC CTG GGC CTC CCT TAG GGA TCC-3'
   NdeI                                                        BamHI
(SEQ ID NO:5)
```

pACTSfiI:

10
```
          SfiI                   SfiI
5'-GAG GCC GAG GTG GCC TAG GGC CTC CTG GGC CTC TAG AAT TCC-3'
                                                  EcoRI
(SEQ ID NO:6)
```

SfiI sites were introduced at the beginning and end
of *H-RAS* genes by use of PCR and oligonucleotides such that
when cloned in pASSfiI the *RAS* coding region was in frame
15 with the GAL4 DNA-Binding Domain, thus creating a fusion
protein RAS-GBD.   In an identical manner VEGF was cloned into
pASSfiI.   A RAF fusion gene with the GAL4 Activation Domain
was constructed and cloned into pACTSfiI to create RAF-GAD.
Similarly VEGF and KDR were also cloned into pACTSfiI.   The
20 oligonucleotides used for amplification of RAS were as
follows:

5'-G GAC TAG GCC GAG GTG GCC GGT ATG ACG GAA TAT AAG CTG GTG-
3' (SEQ ID NO:7)

5'-G GAC TAG GCC GAG GTG GCC GGA GAG CAC ACA CTT GCA GCT-3'
25 (SEQ ID NO:8)

The oligonucleotides used for amplification of RAF were as
follows:

5'-G GAC TAG GCC GAG GTG GCC ATG GAG CAC ATA CAG GGA GCT-3'
(SEQ ID NO:9)

30 5'-G GAC TAG GCC GAG GTG GCC CGA CCT CTG CCT CTG GGA GAG-3'
(SEQ ID NO:10)

The oligonucleotides used for amplification of VEGF were as
follows:

5'-G GAC TAG GCC GAG GTG GCC GGA GGA GGG CAG AAT CAT CAC-3'
35 (SEQ ID NO:11)

5'-G GAC TAG GCC TCC TGG GCC ACG CCT CGG CTT GTC ACA TCT GC-
3' (SEQ ID NO:12)

The oligonucleotides used for amplification of KDR were as follows:

5'-G GAC TAG GCC GAG GTG GCC CTC TCT GTG GGT TTG CCT AGT GTT TC-3' (SEQ ID NO:13)

5 5'-G GAC TAG GCC TCC TGG GCC CTC CTT TGA AAT GGG ATT GGT AAG-3' (SEQ ID NO:14)

        The reporter strains YULH and N106' were transformed with each of the plasmids containing the fusion genes (RAS-GBD, RAF-GAD, VEGF-GBD, VEGF-GAD, KDR-GAD etc.) to

10 yield YULH(RAS-GBD), YULH(VEGF-GBD), N106'(RAF-GAD), N106'(VEGF-GAD), and N106'(KDR-GAD). When two are mated together (e.g., YULH(RAS-GBD) x N106'(RAF-GAD)), then the interaction between RAS-GBD and RAF-GAD reconstitutes the GAL4 transcription factor, thus activating the *URA3*, *HIS3* and

15 the *lacZ* reporter genes which are under the control of the *GAL* promoter.

6.5.   CONSTRUCTION OF cDNA LIBRARIES IN pASSfiI

20   Following cDNA synthesis from human placental tissue as described above in Section 6.1.6, SfiI adapters were ligated to the cDNA under standard linker ligation conditions. The SfiI adapters used for linker ligation have the sequence:

25   5'- A G G C C G G A G G C-3' (SEQ ID NO:15)
                   5'- T C C T C C G G C C T C C G-3' (SEQ ID NO:16)

        The SfiI linked cDNA was amplified by a PCR of 20 cycles and the primer used in the amplification was:
5'-AGGTGCAAGGCCCAGGAGGCCGGAGGC-3' (SEQ ID NO:17)

30   The first 5 cycles of PCR had the following profile:
94°C for 30 sec
37°C for 30 sec
72°C for 30 sec

35   The next 15 cycles of PCR had the following profile:
94°C for 30 sec

65°C for 30 sec

72°C for 30 sec

The amplified cDNA was digested with SfiI and cloned into pASSfiI that had been digested with SfiI. This
5 created a cDNA library with cDNA fused to the DNA-binding domain of GAL4. A library was of 2.5×10^4 clones was made by this method. The library was propagated in the *E. coli* strain XL1Blue (Stratagene) and plasmid DNA was extracted by standard procedures.
10

6.6.   TRANSFORMATION OF THE REPORTER
       STRAINS WITH THE pASSfiI AND
       pACT cDNA LIBRARIES TO CREATE
       "M" AND "N" POPULATIONS

Plasmid pACT differs from pACT2 in the polylinker
15 region (Durfee et al., 1993, Genes Dev. 7:555-569). The strains YULH and N106' were transformed with the pASSfiI and pACT cDNA libraries by the lithium acetate protocol (Section 6.1.2; Ito et al., 1983, J. Bacteriol. 153:163-168). 1 μg of library DNA generally yields a maximum of about 1×10^6
20 transformants. The pACT cDNA library (gift of Stephen J. Elledge; Baylor College of Medicine) (Durfee et al., 1993, Genes Dev. 7:555-569) consists of human peripheral T lymphocyte cDNA and the pASSfiI cDNA library consists of human placental cDNA as described in Section 6.5. The
25 transformants were selected on either media lacking leucine (for pACT) or lacking tryptophan and containing 5-FOA (for pASSfiI). In the latter case, all GBD-fusions that fortuitously activate transcription on their own are eliminated as the URA+ cells will be killed. The
30 transformants were harvested in the appropriate media (SC-LEU for pACT and SC-TRP for pASSfiI) to a final cell density of 1 $\times 10^8$ cells/ ml and stored in aliquots at -70°C after making them 10% in DMSO or glycerol.

35

6.7. CONSTRUCTION OF YEAST STRAINS
WITH INTEGRATED COPIES OF RAF-GAD

As an alternative to bearing RAF on a replicating
plasmid, the RAF-GAD fusion gene was integrated into the
5 yeast genome. This method has the advantage of creating
stable strains of yeast that express RAF-GAD which do not
require growth on selective media for the maintenance of the
fusion gene.

The RAF-GAD fusion gene was amplified by PCR from
10 the RAF-GAD plasmid, using the following oligonucleotides:

          Eag I
5'-GGG ACA AAC GGC CGC ACC GAA ACG CGC GAG GCA GCA AC-3' (SEQ
ID NO:18)

          Sph I
5'-GGG AGT TGC ATG CGC CGG TAG AGG TGT GGT CAA TAA G-3' (SEQ
15 ID NO:19)

These oligonucleotides also introduced unique
restriction sites (EagI and SphI) that facilitate the cloning
of the amplified DNA fragments into the integration vector
20 R1400. The R1400 plasmid vector consists of two yeast genes
LYS2 (Fleig et al., 1986, Gene 46:237-245) and *MER2*
(Engebrecht et al., 1991, Cell 66:1257-1268). The *LYS2*
marker is used for the selection of integration events, while
the *MER2* gene is used for integration of the entire plasmid
25 into the yeast genome. *MER2* is a gene that is not essential
for the vegetative growth of yeast. The RAF-GAD gene was
cloned into the R1400 plasmid to yield RAF-INT. This plasmid
was then digested with the restriction enzyme PstI that has a
site in the *MER2* gene. The restriction was done in a partial
30 manner as there are other PstI sites in the plasmid vector.
The restriction digestion was allowed to proceed only for 1
minute and then the enzyme was inactivated by extracting with
phenol-chloroform and the DNA was then precipitated. This
linearized DNA was used to transform the YULH strain to yield
35 YULH-RAFINT. Integration occurs at the *MER2* locus, and the
integration events were selected by growing the transformants
on media lacking lysine. The N106' strain was transformed

with the RAS-GBD plasmid to form N106' (RAS-GBD). The proper
functioning of the RAF-GAD fusion was confirmed by mating
YULH-RAFINT and N106' (RAS-GBD) and observing the resultant
activation of the *URA3* and the *lacZ* Reporter Genes.

5

6.8. CONSTRUCTION OF PEPTIDE EXPRESSION VECTORS (PEVs)

The PEVs serve to express and localize peptides (or
proteins) in the nucleus of the yeast cell where their
10 potential to inhibit specific protein-protein interactions is
tested. This inhibitory activity of the peptides is
monitored by their ability to inhibit the activity of the
reporter genes like *URA3*, *HIS3* and *lacZ*.

The PEVs comprise the following operably linked
15 components (Figure 7): an *ADC1* promoter (*ADC1-P*) for
supporting transcription in yeast (Ammerer, 1983, Meth.
Enymol. 101:192-201); a nucleotide sequence encoding an SV40
Nuclear Localization Signal (NLS) for transporting the
peptide to the nucleus (Dingwal and Laskey, 1991, Trends.
20 Biochem. Sci. 16:478); followed by a stop codon (UAG) for
terminating translation; means for inserting a DNA sequence
encoding a candidate inhibitor peptide into the PEV in such a
manner that the candidate inhibitor peptide is capable of
being expressed as part of a fusion protein containing the
25 NLS; and an ADC1 transcription termination signal. The NLS
from SV40 large T comprises a 7 amino acid stretch (PKKKRKV)
(SEQ ID NO:20) that has been successfully used in targeting
proteins into the yeast nucleus (Benton et al., 1990, Mol.
Cell. Biol. 10:353-360). The *ADC1* promoter and the sequence
30 encoding the NLS are separated by two restriction sites for
Sfi I and Asc I, respectively, that facilitate cloning of
insert DNAs encoding the peptides. These sites can also be
used for introducing a polypeptide backbone into which the
inhibitory peptide can then be fused; this can facilitate the
35 proper folding and presentation of the peptide. The PEVs
also contain 2μ DNA for replication in yeast, a *LEU2* gene for

selection in yeast, and a β-lactamase gene for selection in
E. coli.

pPEV1 is constructed in the following manner.
Synthetic oligonucleotides that introduce SfiI and AscI
5 sites, the NLS and a stop codon are cloned into the HindIII
site in pAAH5 (Ammerer, 1983, Meth. Enzymol. 101:192-201).
pAAH5 has the ADC1 promoter that supports transcription of
genes in yeast and is marked with LEU2 for selection of
transformants. The sequences of the two oligonucleotides are
10 as follows:

ADCNL8-S

5'-AGC TTG GCC TCC CAG GCC ACA GAC AGG CGC GCC CCC AAA GAA
GAG AAA GGT TTA GA-3' (SEQ ID NO:21)

15 **ADCNL8-A**

5'-AGC TTC TAA ACC TTT CTC TTC TTC TTT GGG GGC GCG CCT GTC
TGT GGC CTG GGA GGC CA-3' (SEQ ID NO:22)

                    6.9. SELECTION OF PROTEIN-PROTEIN
20   INTERACTIONS FROM A
                          NON-INTERACTING BACKGROUND

A. *Selection of SNF1-SNF4 interactions: Mating assay*

SNF1 and SNF4 are a pair of interacting proteins in
the yeast *Saccharomyces cerevisiae* (Celenza and Carlson,
1986, Science 233:1175-1180). The following example
25 describes the selection of these two interacting proteins
SNF1 and SNF4, from a background of cells that do not contain
any DNA-binding or activation domain fusion proteins. This
experiment provides an example of the selection of cells
expressing interacting proteins from a population. The yeast
30 strains expressing these two interacting proteins as fusions
to the DNA-binding and activation domains of GAL4 were mated
in the presence of varying quantities of yeast strains that
were not expressing any fusion protein. As evidenced from
the results below, selection of SNF1-SNF4 interaction occurs
35 even at a 100 to 1000-fold excess of background (cells that
do not contain interacting proteins).

The reporter strains N105 and N106 were transformed with the SNF4-GAD (called pSE1111, a gift from Stephen J. Elledge, Baylor College of Medicine; Fields and Song, 1989, Nature 340:245-246) and SNF1-GBD (called pSE1112, a gift from
5 Stephen J. Elledge, Baylor College of Medicine; Fields and Song, 1989, Nature 340:245-246) to yield N105 (SNF4-GAD) and N106 (SNF1-GBD), respectively.

N105 (SNF4-GAD) and N106 (SNF1-GBD), were grown in the appropriate selective media to a cell density of 1 x 10^8
10 cells per ml. The SNF1-GBD and SNF4-GAD transformants were mixed with the a and α reporter strains, transformed with the vector pAS2 (in N105) and the vector pACT2 (in N106) respectively, in the following dilutions:

2.5 x 10^5 cells of SNF1-GBD and SNF4-GBD strains
15 each mixed with 2.5 x 10^5 cells each of an a strain bearing pAS2 and an α strain bearing pACT2.

2.5 x 10^4 cells of SNF1-GBD and SNF4-GBD strains each mixed with 2.5 x 10^5 cells each of an a strain bearing pAS2 and an α strain bearing pACT2.
20   2.5 x 10^3 cells of SNF1-GBD and SNF4-GBD strains each mixed with 2.5 x 10^5 cells each of an a strain bearing pAS2 and an α strain bearing pACT2.

2.5 x 10^2 cells of SNF1-GBD and SNF4-GBD strains each mixed with 2.5 x 10^5 cells each of an a strain bearing
25 pAS2 and an α strain bearing pACT2.

The mixtures were plated in a volume of 500 μl onto YPAD plates and incubated at 30°C for 8 hours. (During this incubation, one or two cell divisions may occur resulting in duplication of events.) After this, the cells were harvested
30 by the addition of 500 μl of SC-LEU-TRP medium and plated onto media lacking leucine, tryptophan, histidine and containing 40 mM 3-aminotriazole (3-AT).

After three to six days, the number of TRP+, LEU+, HIS+ and 3-AT resistant colonies were counted. Results from
35 our completion of this protocol are shown in Table 11.

Table 11

No. of cells of SNF1-GBD and SNF4-GAD each	No. of cells of pAS2 and pACT2 each	No. of TRP+, LEU+, HIS+, 3-AT^r colonies
2.5×10^5	2.5×10^5	Confluent growth (>10,000 colonies)
2.5×10^4	2.5×10^5	458
2.5×10^3	2.5×10^5	7
2.5×10^2	2.5×10^5	1
0	2.5×10^5	0

Confirmation of interaction by whole cell PCR

Whole cell PCR was performed on the cells positive for interactions as described under the protocols section (Section 6.1.6):

Reaction volume : 100 μl

10X PC2 Buffer for Klentaq : 10 μl

10 mM dNTPs : 3 μl

50 pmoles of each primer pair

1.0 μl of Klentaq polymerase

A few yeast cells from the colony (a swipe of the colony that is positive for interaction with a plastic tip).

PCR was performed at 94°C for 30 sec, 55°C for 30 sec and 72°C for 2 min with each being repeated for 20-30 cycles. Two separate PCR reactions were performed in parallel on the colonies that were TRP+, LEU+, HIS+ and 3-AT resistant. One PCR with the pASFOR (ATGAAGCTACTGTCTTCTATCGAAC-3') (SEQ ID NO:4) and pACTBAC (5'-CTACCAGAATTCGGCATGCCGGTAGAGGTGTGGTCA) (SEQ ID NO:3) primers (for pAS2) amplifies the insert from the GAL4 binding domain fusion (GBD) plasmid, and the other PCR with the pACTFOR (5'-ATGGATGATGTATATAACTATCTATTC-3') (SEQ ID NO:122) and pACTBAC primers (for pACT or pACT2) amplifies the insert from

the GAL4 activation domain fusion (GAD) plasmid. As
controls, PCR reactions were performed on cells that harbored
the GBD and GAD plasmids that did not contain any insert DNA.

"Real positives," in which the pAS2 and pACT2
5 vectors are replaced by analogous vectors containing for
example, cDNA inserts, should yield PCR products for both the
GBD and GAD plasmids that are bigger than that of the
respective controls. pAS2 and pACT2 specific primers are
used in a yeast whole cell PCR assay on these colonies. In a
10 trial, PCR products whose sizes corresponded to SNF1- and
SNF4-fusion proteins were obtained.

B. Selection of VEGF-VEGF interactions: Mating assay

The following example describes the optimum plating
15 conditions for the selection of interacting proteins from a
mating assay. The yeast strains expressing two interacting
proteins, both VEGF in this case, as fusions to the DNA-
binding and activation domains of GAL4, were mated in the
presence of varying quantities of yeast strains that were not
20 expressing any fusion protein. The effect of increasing the
total cell density on the efficiency of selecting the VEGF-
VEGF interaction was studied as described below.

YULH(VEGF-GBD) and N106'(VEGF-GAD), made as
described in Section 6.4, were grown to saturation in media
25 (SC-TRP-LEU) that selects for both of these plasmids in which
VEGF is encoded. VEGF dimerizes to form homodimers (Potgens
et al., 1994, J. Biol. Chem. 269:32879-32885; Claffey et al.,
1995, Biochem. Biophys. Acta 1246:1-9), and thus the
interaction between two VEGF molecules can be monitored in
30 the mating interaction assay. Simultaneously, YULH and N106'
were grown to saturation in YPAD medium. The VEGF-GBD and
VEGF-GAD transformants were mixed with the YULH and N106'
reporter strains, in the following dilutions:

6.6 x 10^4 cells of YULH(VEGF-GBD) and N106'(VEGF-
35 GBD) strains each mixed with 6.6 x 10^7 cells each of YULH and
N106' strain in a total volume of 0.5 ml.

1.3 x 10^5 cells of YULH(VEGF-GBD) and N106'(VEGF-
GBD) strains each mixed with 1.3 x 10^8 cells of YULH and N106'
strain in a total volume of 1 ml.

2.6 x 10^5 cells of YULH(VEGF-GBD) and N106'(VEGF-
5 GBD) strains each mixed with 2.6 x 10^9 cells each of YULH and
N106' strain in a total volume of 2 ml.

5.3 x 10^5 cells of YULH(VEGF-GBD) and N106'(VEGF-
GBD) strains each mixed with 5.3 x 10^8 cells of YULH and N106'
strain in a total volume of 4 ml.

10   The mixtures were transferred onto one YPAD plate
each, each plate 150 mm in diameter, and incubated at 30°C
for 6-8 hours.   (During this incubation one or two cell
divisions may occur resulting in duplication of events).
After this, the cells were harvested by the addition of 1-2
15 ml of SC-LEU-TRP-URA-HIS medium and plated onto plates
lacking leucine, tryptohan, histidine, uracil and containing
40 mM 3-aminotriazole (3-AT). The contents of one YPAD plate
went into one selective media plate.

After three-six days, the number of TRP+, LEU+,
20 HIS+, URA+ and 3-AT resistant colonies were counted. In an
exemplary trial, the following results shown in Table 12 were
obtained:

Table 12

25

No. of YULH and N106 cells each	No. of VEGF (GBD) and VEGF (GAD) cells each	No. of HIS +, URA + and 3-AT' colonies[1]
6.6 x 10^7	6.6 x 10^4	71
1.3 x 10^8	1.3 x 10^5	137
2.6 x 10^8	2.6 x 10^5	233
5.3 x 10^8	5.3 x 10^5	*

*   The paste representing the mixture of cells was so thick that the emerging colonies could not be clearly differentiated from the background.

35   [1]   These values represent averages of duplicates.

VEGF-VEGF interactions were detected. The optimum
cell density required for mating to yield interacting
colonies was 1-2 x 10⁸ cells/150 mm diameter plate, since at
cell densities higher than this, the number of interactants
5 detected decreased. At cell densities higher than 1-2 x 10⁸
cells/plate, doubling the individual interacting cells did
not double the yield of HIS⁻, URA⁻ and 3-AT^R cells.

6.10. SELECTION OF SPECIFIC PROTEIN-PROTEIN INTERACTIONS FROM A BACKGROUND OF OTHER INTERACTING PROTEINS

10

Mating assay

This example describes the selection of a pair of
interacting proteins from a background of other interacting
proteins. The interaction between the RAS-GAL4 DNA binding
15 domain fusion and RAF-GAL4 activation domain fusion proteins
was selected in the presence of other GAL4 DNA-binding and
activation domain fusion proteins. This example demonstrates
that specific interactors can be selected when present in a
background of other interacting proteins.
20
YULH(RAS-GBD) and N106'(RAF-GAD) transformants made
as described in Section 6.4 were grown in the appropriate
selective media to a cell density of 2 x 10⁸ cells/ ml. RAS
and RAF are members of signal transduction pathway leading to
mitogenesis and have been demonstrated to interact with each
25 other (Vojtek et al., 1993, Cell 74:205-214). The RAS-GBD
and RAF-GAD transformants were mixed with the M and N cells
in the following dilutions:

2.5 x 10⁵ cells of RAS-GBD and RAF-GBD strains each
mixed with 2.9 x 10⁸ cells each of M and N.
30
2.9 x 10⁸ cells each of M and N.

The 'M' cells in this example are YULH cells
bearing a library of human placental cDNA fused to GBD in
pASSfi. The 'N' cells in this example are N106' cells
bearing a library of cDNA of human peripheral T lymphocytes
35 fused to GAD in pACT.

The M and N cells represent 1000 transformants
each. That is, in 10^8 cells each transformant is represented
10^5 times.

These mixtures were vortexed very gently and then
5 pelleted by centrifugation and then resuspended in 0.2 ml of
YPAD medium and then spread on YPAD plates and incubated at
30°C for 6-8 hours. (During this incubation one or two cell
divisions may occur resulting in duplication of events.) At
this stage, cells of the M (and RAS-GBD) and N (and RAF-GAD)
10 strains have mated to yield diploids.

The mating mixes were harvested from the plates by
adding 1 ml of SC-URA-LEU-TRP media and scraping. The
harvested cells were then plated onto SC-URA-LEU-TRP-HIS+3-AT
agar plates. The -TRP and -LEU select for the GBD and GAD
15 plasmids (encoding trp and leu, respectively), while the -URA
and -HIS and presence of 3-AT selects for the interaction
between the two fusion proteins (by selecting for the
expression of the *URA3* and *HIS3* Reporter Genes). Thus, cells
that are URA+, HIS+, 3-AT resistant, TRP+ and LEU+ contain
20 GAD and GBD fusion proteins that interact with each other.

The URA+ cells were picked and patched onto SC-LEU-
TRP plates and incubated at 30°C for 12-24 hours. These
patches were then replica-plated onto SC-URA, SC-HIS and SC-
TRP-LEU plates. Growth on -URA and -HIS plates confirmed
25 interaction of the two fusion proteins. The patches from the
SC-LEU-TRP plates were transferred onto a Whatman no. 1
filter and assayed for β-galactosidase activity (Section
6.1.11). The patches turned blue, indicating β-galactosidase
activity as a result of the activation of the *lacZ* Reporter
30 Gene due to interaction between RAS-GBD and RAF-GAD.

Confirmation of interaction by whole cell PCR

Two PCR reactions were performed in parallel on the
colonies that were TRP+, LEU+, and URA+ (as in the case of
35 Section 6.3): one with the RAFSfiS (5'-G GAC TAG GCC GAG GTG
GCC GGT ATG ACG AAA TAT AAG CTG GTG-3') (SEQ ID NO:23) and
RAFSfiA (5'-G GAC TAG GCC GAG GTG GCC GGA GAG CAC ACA CTT GCA

- 273 -

GCT-3') (SEQ ID NO:24) that amplify the RAF insert from the GAD plasmid, and the other with primers specific for the RAS sequences, RASSfiS (5'-G GAC TAG GCC GAG GTG GCC ATG GAG CAC ATA CAG GGA GCT-3') (SEQ ID NO:25) and RASSfiA (5'-G GAC TAG

5 GCC GAG GTG GCC CGA CCT CTG CCT CTG GGA GAG-3') (SEQ ID NO:26).

RAF-RAS interactors yield PCR products for the GAD plasmid with the RAF specific primers and a RAS-specific PCR product with the RAS specific primers. The ratios of the

10 RAS-RAF interactors to the total cells in each mating, shown in Table 13, were obtained:

<div align="center">Table 13</div>

No. of RAS-GBD and RAF-GAD cells each in the mating mix	No. of M and N cells each in the mating mix	Total no. of RAS-RAF interactants
0	2.9×10^6	0
2.5×10^5	2.9×10^8	200'

' This value represents average of duplicates.

6.11. SELECTION OF INTERACTING PROTEINS FROM AN M x N SCREEN

6.11.1. MATING ASSAY

The M and N cells (as described in Section 6.10) were mixed together and 0.5 ml of the mix (a total cell density of 2.5×10^9 cells /ml) was spread onto YPAD plates and incubated at 30°C for 8 hours for mating. The M and N cells represent 5,000 transformants each. That is, in 10^8 cells each transformant is represented 20,000 times. The mating mixes were then harvested in 1 ml of the appropriate selective media and plated onto SC-URA-LEU-TRP-HIS plates that contain 40 mM 3-AT and incubated at 30°C until colonies emerge. In a trial, this analysis was performed in duplicate.

Cells that were URA+, HIS+ and 3-AT resistant were
patched onto separate SC-TRP-LEU plates and assayed for
β-galactosidase activity. Cells that were URA+, HIS+, 3-AT
resistant and positive for β-galactosidase activity were
5 classified as positive for protein-protein interactions
(Sections 6.1.8 and 6.1.10). These colonies were then grown
to saturation in 100 μl each of SC-LEU-TRP medium in a 96
well plate and an aliquot was stored frozen after making it
10% in DMSO. These cultures represent the interactive
10 population from an M x N screen.

6.11.2. <u>WHOLE CELL PCR OF THE POSITIVE COLONIES</u>

From the patches of the positive colonies, whole
cell PCR was performed as described under Section 6.1.8 with
15 the modification that a tiny amount of the colony was taken
with the help of a plastic tip and transferred to the PCR mix
for amplification of the inserts from the GBD (in pASSfiI)
and GAD (in pACT) plasmids. Two PCR reactions are performed
in parallel for each colony: one with the pASFOR (SEQ ID
20 NO:4) and pACTBAC (SEQ ID NO:3) primers that amplify the
insert from the GBD plasmid, and one with the pACTFORII (SEQ
ID NO:2) and pACTBAC (SEQ ID NO:3) primers that amplify the
insert from the GAD plasmid.

The primers can be used for sequencing as well as
25 PCR.

6.11.3. <u>QEA™ METHOD OF THE PCR PRODUCTS</u>

The pASSfiI and pACT specific PCR products were
pooled separately and a 4-mer and 5-mer QEA™ method were
30 performed as described in in Section 6.1.12.2.1. 10 μl of
each PCR reaction were used in pooling. The pooled PCR
products were then purified with the GeneClean II DNA
purification kit (Bios 101) according to the manufacturer's
instructions. The GeneClean II kit uses a glassmilk-based
35 DNA purification protocol. 10 ng of the pooled PCR products
were used in a QEA™ method reaction. The enzymes Sau3A I and

- 275 -

BsaW I were used in the QEA™ method process.   The primer
pairs for QEA™ method were as follows:

For Sau3A I,   5'-AGCACTCTCCAGCCTCTCACCGAC-3' (SEQ ID NO:27)
 5   3'-AGTGGCTGCTAG-5' (SEQ ID
                                          NO:28)

For BsaW I,   5'-AGCACTCTCCAGCCTCTCACCGAC-3' (SEQ ID NO:29)
                           3'-AGTGGCTGGGCC-5' (SEQ ID
                                          NO:30)

   The QEA™ method products were then analyzed on a ABI 377
10 automated sequencer by denaturing gel electrophoresis.   In a
   trial, the QEA™ method patterns obtained from duplicate N x M
   screens were very similar (Figure 8).

6.11.4.   CREATION OF TWO-DIMENSIONAL POOLS

15   Two-dimensional pools were created as per Section
   6.1.9.   5 µl of saturated culture from each well in a row or
   in a column were combined to create a pool which was given a
   particular designation (like Pool 1, 2, 3... for columns and
   Pool A, B, C... for rows).   Each of these pools then served
20 as starting material for further analysis by PCR.   A
   duplicate of the two-dimensional pool was made in which an
   additional well, that consisted of diploids resulting from
   the mating of YULH(RAS-GBD) and N106'(RAF-GAD), was added to
   this array.
25

6.11.5.   WHOLE CELL PCR OF THE POOLED CELLS

            Whole cell PCR was performed on the pooled rows and
   columns arising from the two-dimensional pools as described
   under the protocols section (Section 6.1.8).   Two PCR
30 reactions were performed in parallel for each pool: one with
   the pASFOR (SEQ ID NO:4) and pACTBAC (SEQ ID NO:3) primers
   that amplify the insert from the GBD plasmid, and one with
   the pACTFORII (SEQ ID NO:2) and pACTSEQII (SEQ ID NO:1)
   primers that amplify the insert from the GAD plasmid.   Thus,
35 each PCR reaction represents genes from a particular pool for
   either the "M" or the "N" population.   The PCR products

served as templates for further analysis by the QEA™ method
and SEQ-QEA™ method.

6.11.6. QEA™ METHOD OF THE PCR DERIVED FROM POOLED CULTURES

PCR products (10 μl out of 100 μl) from each row or
column (in the case of two-dimensional pools) were all
combined and subjected to the QEA™ method as described above.
The QEA™ method was also performed on the PCR products from
the individual rows and columns. Four base-pair recognition
site restriction enzymes like Sau3A I, BsaW I and Tsp 509 I
were used and after restriction-digestion for 120 min, the
enzymes were either heat-inactivated at 65°C for 20 min or
inactivated by phenol extraction. Combinations of four-base
recognizing enzymes (Sau3A I) and six-base recognition
enzymes (Hind III) were also used in the QEA™ method.

For Tsp 509 I, the QEA™ method primer pair
(adapter) used in a trial was:

    5'-AGCACTCTCCAGCCTCTCACCGAC-3' (SEQ ID NO:31)
                    3'-AGTGGCTGAATT-5' (SEQ ID NO:32)

For Hind III, the QEA™ method primer pair (adapter)
used in a trial was:

    5'-AGCACTCTCCAGCCTCTCACCGAC-3' (SEQ ID NO:33)
                    3'-AGTGGCTGTCGA-5' (SEQ ID NO:34)

SEQ ID NO:31 and NO:33 had the fluorescent dye Fam affixed to
the 5' end.

For Sau3AI, the QEA™ method primer pair (adapter)
used in a trial was

    Primer RC24: 5'-AGCACTCTCCAGCCTCTCACCGAC-3' (SEQ ID
NO:67)   3'-AGTGGCTGCTAG-5'
(SEQ ID NO:74):   Primer RC9.

Primer RC24 had biotin attached at its 5' end.
After this, T4 DNA ligase was added and the QEA™ method was
performed as described in Section 6.1.12.2.1.

The QEA™ method was carried out with Sau3AI and
HindIII, using the above primer pairs listed for each enzyme.
The QEA™ method products were analyzed on denaturing

polyacrylamide gels as described above. Each QEA™ method
band is a representative of protein present in the
interactive population. The QEA™ method patterns obtained
from duplicate N x M screening trials were very similar.

5 This was observed with both pAS-and pACT-specific PCR
products. A RAF-specific band was clearly seen, at the
expected position, in the QEA™ method of the pool that
contained the RAS-RAF diploid, while this band was absent in
the pool that did not contain the RAS-RAF diploid (Figure 9).

10   Furthermore, by comparing the QEA™ method patterns
of each row and column, it was possible to identify the well
from which the RAS-RAF diploid originated. This is exemplary
of deconvolution of the QEA™ method results from the two-
dimensional pool to arrive at source of genes that contribute

15 to differential QEA™ method patterns.

        Thus, by analyzing multiple replicates of the QEA™
method of one interactive population with multiple replicates
of the QEA™ method of another interactive population, it is
possible to identify genuine differences and thus to identify

20 and isolate protein-protein interactions that are unique to
any particular tissue/cell type, stage of development, or
disease state.

            6.11.7.   THE SEQ-QEA™ METHOD OF THE PCR
25   DERIVED FROM POOLED CULTURES

        The QEA™ method products from the pooled pASSfiI
and pACT PCR products are subjected to the SEQ-QEA™ method
separately as described in Section 6.1.12.2.2. The SEQ-QEA™
method gives the additional information about each QEA™-

30 method-product in that it provides the identity of the
terminal 4 bases immediately downstream of the restriction
site that is used in the QEA™ method. With this additional
information, gene identification is possible even with 4-base
recognizing restriction enzymes. Comparison of the QEA™

35 method and the SEQ-QEA™ method patterns between the rows and
columns of the pooled interactants (see Figure 3) permits the
deconvolution of the pools and thus reveals the location of

each interacting pair in the original master plate that
contains all the interacting pairs. Gene identification
through the SEQ-QEA™ method thus reveals the identity of each
pair of interacting proteins in an interacting population and
5 thus helps in the identification of unique interactors
specific to a particular population.

6.12. IDENTIFICATION OF SPECIFIC PAIRS OF INTERACTING PROTEINS FROM A QEA™ METHOD OF THE INTERACTIVE POPULATION AND BY THE USE OF GENE-SPECIFIC PRIMERS

10

Based on the sequence information obtained from the
SEQ-QEA™ method, gene-specific primers are synthesized and
are used as PCR primers to screen the interactive population.
PCR is performed on all the pools of PCR products (that are
15 derived from the interactive population from the pAS-like
vectors and from the pACT-like vectors using the gene
specific primers; see Section 6.1.8). Decoding the PCR
results identifies the original colony that gives rise to the
QEA™ method band. The pAS-like-vector and pACT-like-vector
20 primer-derived PCR products from these colonies are then
sequenced to reveal the identities of both the interacting
proteins. The identity of one of the genes encoding the
interacting proteins is given by the sequence obtained from
the QEA™ method band.
25

6.13. CREATION OF INTERACTIVE GRIDS

As an alternative to the above PCR-based strategy
to identify interacting proteins from an interactive
population, a hybridization-based strategy is used. As a
30 first step in this process an "Interactive Grid" is created
in the following manner. A portion (25 μl) of the pooled PCR
products (derived using the pAS-like-vector-specific and
pACT-like-vector-specific primer pairs) are used to create an
interactive grid. The interaction grids are created by
35 spotting a pair of PCR products onto a nylon membrane with
the same dimensions as the 96-well plate from which the
whole-cell PCR was done. The DNA is denatured according to

standard protocols before spotting onto a nylon membrane.
Spotting of DNA is done as per standard dot-blotting
protocols for RNA, except with prior denaturation (in Current
Protocols in Molecular Biology, 1995, Chapter 2.9B, Dot and
5 Slot Blotting of DNA onto uncharged nylon and nitrocellulose
membranes, Frederick M. Ausubel et al. (eds.), John Wiley &
Sons, New York). Thus, each spot on the interactive grid
corresponds to the original well containing the culture
harboring the two interacting proteins.
10

6.14.   ISOLATION OF STAGE-SPECIFIC
PAIRS OF INTERACTING PROTEINS

The QEA™ method stage/tissue-specific bands are
excised from gels and amplified by PCR using the same primer
15 sets that are used in the QEA™ method. These PCR products
are then labeled either with radiolabeled nucleotides (e.g.,
^{32}P-dCTP) or biotinylated nucleotides (e.g., Bio-dCTP) or
fluorescently tagged nucleotides, and used to probe the
interaction grids. Labeling and hybridization are done
20 according to standard protocols (Sambrock et al., 1989,
Molecular Cloning: A Laboratory Manual, Second Edition, Cold
Spring Harbor Laboratory Press, Cold Spring Harbor, New
York). Spots that hybridize to the probe represent the pair
of interacting proteins from which the QEA™ method band
25 arose. By relating this signal to the original master plate,
the original cell culture harboring the two interacting
proteins can be identified.

To sequence the pAS-like vector and pACT-like
vector clones from these cells, the stored PCR products (50
30 μl each) are sequenced by standard protocols and the sequence
identity is obtained.

6.15.   EXPRESSION OF PEPTIDE INHIBITORS
IN PEV AND INHIBITION OF
PROTEIN-PROTEIN INTERACTIONS

35 To test the functionality of pPEV1 (described in
Section 6.8), the RAS effector peptide (amino acids 17-40) is
cloned between the Sfi I and Asc I sites to yield pPEVRAS-E,

- 280 -

which is used to transform the yeast strain YULH-RAFINT by
the lithium acetate protocol (Section 6.1.2). The RAS
effector peptide arises from a region in the RAS protein that
is important for its interaction with RAF (Chuang et al.,
5 1994, Mol. Cell. Biol. 14:5318-5325; Zhang et al., 1993,
Nature 384:308-313). The resulting strain, YULH2RAS, is then
mated to the N106'(RAS-GBD) strain. The mated cells are
transferred to the appropriate selective media. The
resulting diploids are ura- (unable to grow on media lacking
10 uracil) and lacz- (negative for β-galactosidase activity).
These diploids also grow on medium containing 5-FOA, a
chemical that kills URA+ cells (Rothstein, 1983, Meth.
Enzymol. 101:167-180). The control N106'(RAS-GBD) diploids
are URA+ and LACZ+, but are unable to grow on medium
15 containing 5-FOA.

Thus, pPEV1 can be successfully used to introduce a
polypeptide into the nucleus where this polypeptide
successfully competes with and inhibits a specific protein-
protein interaction. In the above instance the RAS-2 peptide
20 inhibits the interaction between the RAF-GAD and RAS-GBD
proteins. Furthermore, the presence of the inhibitory
peptide enables the cells to grow in the presence of an agent
(like 5-FOA) that would kill or select against cells
displaying interaction between the two proteins. Thus, this
25 general method has use as a device to screen for and isolate
peptides or other inhibitors that can specifically inhibit
protein-protein interactions.

6.16. IDENTIFICATION OF CELLS CONTAINING AN INHIBITOR OF PROTEIN-PROTEIN INTERACTION USING THE 5-FOA ASSAY

30

The above method described for the isolation of
inhibitor peptides can also be used to screen and isolate
inhibitors that are not genetically encoded. In other
contexts, for example, a yeast-based transcription-inhibition
35 assay has been used to screen for inhibitors of the HIV-1
proteinase (Murray et al., 1993, Gene 134:123-128). The

reporter strains YULH(R4-GBD) and N106' (FKBP-12-GAD), that
harbor the reporter gene-systems and the two interacting
proteins are grown in a 96-well format with each well
containing 200 μl of media that selects for both the GBD and
5 GAD plasmids (SC-LEU-TRP).

After a growth period (24-36 hours) that is
sufficient for early log-phase growth (a cell density of 1 x
10^7 cells/ml), the cells are exposed to an RNA inhibitor at a
concentration of 100 nM for 1-2 hours. This RNA inhibitor is
10 of the sequence 5'-CCCUGAUGGUAGACCGGGGUG-3' (SEQ ID NO:35).
The pyrimidines in this RNA are modified with 2'-amino-2-
deoxypyrimidines that causes the RNA to be resistant to
nucleases.   This RNA binds with high affinity to VEGF (Green
et al., 1995, Current Biol. 2:683-695). After treating the
15 cells with the RNA inhibitor, a 1:10 dilution of the cells is
transferred to a 96-well plate containing 200 μl media same
as above except that it lacks uracil, and incubated for 4-6
hours. This medium requires that cells express the URA3 gene
product. As the expression of the URA3 gene is dependent on
20 the interaction between the two hybrid proteins, only those
cells where inhibition is not occurring will express the URA3
gene product. In other words, cells where inhibition occurs
do not express the URA3 gene and hence are ura minus.

After treating the cells with the RNA inhibitor in
25 a medium that lacks uracil, a 1:10 dilution of the cells is
transferred to a 96-well plate containing 200 μl media same
as above except that it contains 5-FOA, uracil and the RNA at
a concentration of 100 nanomolar. FOA kills the URA+ cells
(i.e., cells in which inhibition did not occur); uracil
30 allows the ura minus cells to grow back (i.e., cells where
inhibition occurred), and the presence of RNA inhibitor
ensures that there is no reversion of inhibition.

Growth is evident only in those instances where the
RNA inhibitor is present. The cells are able to grow in the
35 absence of 5-FOA but in the presence of the inhibitor in SC-
TRP-LEU, indicating that absence of growth in 5-FOA is due to
inhibition. In the absence of the RNA inhibitor, cells are

not able to grow in 5-FOA. The activity of the *lacZ* reporter
genes is also assayed enzymatically. The activity of the
lacZ gene is determined by assaying the β-galactosidase
activity of the cells as mentioned in Section 6.1.11. Thus,
5 by selecting for growth in a inhibitor-dependent fashion, a
robust and high throughput assay for the selection of
inhibitor drugs that inhibit protein-protein interactions is
achieved.

10   6.17.   5-FOA INHIBITION ASSAY FOR SELECTING
               INHIBITION OF THE INTERACTION
               BETWEEN R4 AND FKBP-12

      6.17.1.   DEVELOPMENT OF EXPERIMENTAL PARAMETERS
Construction of R4-GBD and FKBP-12-GAD fusion genes:

      The cytoplasmic domain of R4 (also known as ALK5; Frazen
15 et al. 1993, Cell, 75:681) is a Type I receptor for the
Transforming Growth Factor β (TGFβ) that has been
demonstrated to bind to the immunophilin FKBP-12 (Standaert
et al. 1990, Nature 346:671) in the yeast two-hybrid assay
(Wang et al. 1994, Science 265:674-676; Wang et al. 1996,
20 Science 271:1120-1123). This interaction is blocked by the
immunosuppressant drug FK506 in the yeast two-hybrid assay
(Wang et al. 1994, Science 265:674-676; Wang et al. 1996,
Science, 271:1120-1123).

      The interaction between R4 and FKBP-12 is monitored
25 according to the invention by the ability to activate the
lacZ Reporter Gene, and the inhibition of the interaction by
FK506 is monitored by a reduction in the activity of the *lacZ*
Reporter Gene in the presence of FK506.

      The DNA encoding the cytoplasmic domain of R4 was
30 obtained by PCR amplification using total peripheral
T-lymphocyte cDNA as template. The primers used for
amplification were:
ALK5SfiI-S
5'-GGACTAGGCCGAGGTGGCCTGCCACAACCGCACTGTCATTCAC-3'
35 (SEQ ID NO:45)
ALK5SfiI-A

5'-GGACTAGGCCTCCTGGGCCTTACATTTTGATGCCTTCCTGTTGACTGAG-3'
(SEQ ID NO:46)

        These primers flank the region from amino acid 148 to
5 carboxyl terminus of the protein (Frazen et al., 1993, Cell
   75:681). The PCR products were digested with Sfi I and
   cloned at the Sfi I site in pASSfiI to yield R4-GBD, wherein
   the R4 cytoplasmic domain is fused in frame to the
   DNA-binding domain of GAL4.
10   The FKBP gene was amplified from total peripheral
   T lymphocyte cDNA by PCR using the following primers:

   FKBPSfiI-S
   5'-GGACTAGGCCGAGGTGGCCATGGGAGTGCAGGTGGAAACCATC-3'
15 (SEQ ID NO:47)

   FKBPSfiI-A
   5'-GGACTAGGCCTCCTGGGCCTCATTCCAGTTTTAGAAGCTCCAC-3'
   (SEQ ID NC:48)
20   These primers flank the entire coding region of the
   FKBP-12 protein (Standaert et al., 1990, Nature 346:671). The
   PCR products were digested with Sfi I and cloned at the Sfi I
   site in pACTSfiI to yield FKBB-12-GAD, wherein the FKBP-12
   protein is fused in frame to the activation domain of GAL4.
25

                6.17.2. INHIBITION OF R4-FKBP-12
                         INTERACTION BY FK506 AND THE
                         SELECTION OF THESE INHIBITION
                         EVENTS USING THE 5-FOA ASSAY

   Step I. Interaction of R4-GBD with FKBP-GAD:
30   The R4-GBD and FKBP-12-GAD plasmids were
   transformed into the yeast strains YULH and N106',
   respectively, to yield YULH(R4-GBD) and N106'(FKBP-12-GAD).
   These strains are then mated as described in the mating
   protocol (Section 6.1.1). The resulting diploids are patched
35 onto SC-URA-TRP-LEU-HIS+3-AT media. This media is selective
   for the interaction between the two fusion proteins. Growth

                              - 284 -

in this media demonstrates the interaction between the R4-GBD
and FKBP-12-GAD fusion proteins.

Step II. Growth of R4-GBD::FKBP-12-GAD diploids in
5 *non-inducing media:*

      The R4-GBD::FKBP-12-GAD diplopids are inoculated
into media that contains instead of glucose, a carbon source
like lactate that does not induce the expression from the ADH
promoter (Denis et al., 1983, J. Biol. Chem. 258:1165) that
10 is driving transcription of the two fusion genes. The medium
also lacks tryptophan and leucine to maintain the two
plasmids R4-GBD and FKBP-12-GAD. This is repeated in the
presence or absence of FK506 at a final concentration of
1 μM. This concentration of FK506 has been demonstrated to
15 inhibit the interaction of R4 with FKBP-12 in the yeast
two-hybrid system (Wang et al., 1994, Science, 265:674-676).
These cells may or may not be mixed with the
VEGF-GBD::VEGF-GAD diploids (described in Section 6.9.B).

      The different experiments are summarized below in Table
20 14.

Table 14

Experiment	R4-GBD: FKBP-12-GAD	VEGF-GBD:: VEGF-GAD	FK506 (1μM)	Carbon Source
1	+	+	+	Lactate
2	+	+	-	Lactate
3	+	-	-	Lactate
4	-	+	-	Lactate
5	+	-	+	Lactate
6	-	+	+	Lactate

Step III. Growth of R4-GBD::FKBP-12-GAD diploids in inducing
35 *media:*

The cells are grown in the lactate medium for 24-36
hours and then the cell suspensions corresponding to each
individual experiment are then diluted at 1:100 ratio in
SC-URA-LEU-TRP-HIS+FK506(1μM) liquid media and grown for 8-24
5 hours. The carbon source in this medium is glucose that
supports the induction of transcription from the ADH promoter
(Holland and Holland, 1978, Biochemistry 17:4900). Growth in
all the experiments is monitored by measuring OD_{600}.

10

Table 15

Experiment	R4-GBD: FKBP-12-GAD	VEGF-GBD:: VEGF-GAD	FK506 (1μM)	Carbon Source
1	+	+	+	Glucose
2	+	+	-	Glucose
3	+	.	-	Glucose
4	-	+	-	Glucose
5	+	-	+	Glucose
6	-	+	+	Glucose

Growth in this media should be evident in
Experiments 1, 2, 3, 4 and 6 and should be inhibited only in
Experiment 5 due to the inhibition of the R4-FKBP-12
interaction in the presence of FK506, thereby resulting in
25 the non-activation of the URA3 reporter gene. Growth in
Experiments 1 and 6 should occur due to the interaction of
VEGF-GBD with VEGF-GAD that is not inhibited by FK506.

Step IV. Monitoring inhibition of R4-FKBP12 interaction
30 enzymatically by β-galactosidase assays:
As described above, the cells are allowed to grow
for 8-24 hours (in Step III) after which the β-galactosidase
activity is measured in a fraction of the cells using the
FluoReporter lacZ/Galactosidase Quantitation kit (Molecular
35 Probes) according to the manufacturer's protocols.
Alternatively, chemiluminescent β-galactosidase assays are

- 286 -

performed by using the Galacto-Light and Galacto-Light Plus Chemiluminescent reporter assay system for the detection of β-galactosidase (Tropix, Inc., Bedford, MA). β-galactosidase activity is measured in a fraction of the cells using the
5   FluoReporter lacZ/Galactosidase Quantitation kit (Molecular Probes) according to the manufacturer's protocols and a decrease in β-galactosidase activity should be observed in Experiment 5 (+FK506) in comparison to Experiment 3 (-FK506) of Table 15, above.
10

Step V.   Selecting R4-FKBP inhibition by FK506 using the 5-FOA assay

In parallel, the individual experiments (from Step III) are also diluted in a 1: 100 ratio in SC-LEU-TRP-
15 HIS+FK506(1μM)+5-FOA liquid media and incubated at 30°C for 8-42 hr. The experimental setup is shown in Table 16.

Table 16

Experiment	R4-GBD: FKBP-12-GAD	VEGF-GBD:: VEGF-GAD	5-FOA	FK506 (1μM)	Carbon Source
1	+	+	+	+	Glucose
2	+	+	+	-	Glucose
3	+	-	+	-	Glucose
4	-	+	+	-	Glucose
5	+	-	+	+	Glucose
6	-	+	+	+	Glucose

In this instance, growth should be evident in all the
30 experiments of Table 16 except in experiment 3 where the growth should be inhibited. This is because in experiment 3 the R4-GBD::FKBP-12-GAD interaction activates the *URA3* gene and this event is toxic to yeast in the presence of 5-FOA. β-galactosidase activity is measured in a fraction of the
35 cells using the FluoReporter lacZ/Galactosidase Quantitation kit (Molecular Probes) according to the manufacturer's

protocols and a decrease in β-galactosidase activity should be observed in Experiment 3 in comparison to Experiment 5.

Alternatively, dilutions of the individual treatments are plated on SC-LEU-TRP-HIS+FK506(1μM)+5-FOA

5 plates and after a growth period of 8-48 hours ten colonies from each dilution of each treatment are picked up and whole cell PCR (Section 6.1.8) is performed in parallel with VEGF- (SEQ ID NO. 11 and 12 from Section 6.4) and R4-specific primers (ALK5SfiI S (SEQ ID NO:45) and ALK5SfiI A (SEQ ID

10 NO:46)). In this manner, the selection of either VEGF-VEGF or R4-FKBP diploids is monitored by the presence of the specific PCR product. Experiment 5 (R4-FKBP+FK506) should give rise to greater numbers of colonies than Experiment 3 (R4-FKBP-FK506). From Experiment 1 at lower dilutions

15 predominantly R4 PCR product should be obtained indicating the presence of R4- FKBP diploids, and in the higher dilutions VEGF-specific PCR product should be seen very rarely and the R4-specific PCR product should be almost always obtained.

20   The results should indicate a selection of the R4-FKBP diploids due to the inhibition of their interaction by FK506 and thereby the non-activation of the *URA3* Reporter Gene, allowing the R4-FKBP diploids to survive in the 5-FOA media. On the other hand, the VEGF-VEGF interaction is not

25 inhibited by FK506 and as a result this interaction should activate the *URA3* Reporter Gene and thus the VEGF-VEGF diploids should be killed in the 5-FOA media.

6.17.3.   SELECTION OF R4-GBD::FKBP-12-GAD BY
30   THE 5-FOA ASSAY FROM AN M X N ANALYSIS

Isolation of R4-FKBP interactants in a background of interacting proteins from an M x N analysis:

As described in Section 6.1.7, the strains YULH and N106' are transformed with the pAS2-1 and the pAD-GAL4 or

35 pACT2 cDNA libraries, respectively, by the lithium acetate protocol (Section 6.1.2; Ito et al., 1983, J. Bacteriol. 153: 163-168) to yield M and N populations. 1 μg of library DNA

generally yields a maximum of 1×10^6 transformants. The
transformants are selected on either media lacking leucine
(for pAD-GAL4/pACT2) or lacking tryptophan and containing
5-FOA (for pAS2-1). In the latter case all GBD-fusions that
5 fortuitously activate transcription on their own will be
eliminated as 5-FOA kills the URA+ cells. The transformants
are harvested in the appropriate media (SC-LEU for pAD-
GAL4/pACT2 and SC-TRP for pAS2-1) to a final cell density of
1×10^8 cells/ml.

10   A thousand independent transformants each of M and
N cells are mixed together with YULH(R4-GBD) and N106'
(FKBP-12-GAD) in the following cell to cell ratios: 1.3×10^5
cells each of YULH(R4-GBD) and N106'(FKBP-12-GBD) strains
mixed with 1.3×10^8 cells each of M (YULH with GBD fusions)
15 and N (N106' with GAD fusions) in a total volume of 1 ml.
This is done in duplicate.

          The mixtures are subjected to the mating protocol
described in Section 6.1.1. The mating mixtures are
transferred onto one YPAD plate each, each plate 15 mm in
20 diameter, and incubated at 30°C for 6-8 hours. (During this
incubation one or two cell divisions may occur resulting in
duplication of events). After this, the cells are harvested
by the addition of 1-2 ml of SC-LEU-TRP-URA-HIS medium and
plated onto plates lacking leucine, tryptophan, histidine,
25 uracil and containing 40 mM 3-aminotriazole (3-AT). The
contents of one YPAD plate go into one selective media plate.

          After three-six days the number of TRP+, LEU+,
HIS+, URA+ and 3-AT resistant colonies are picked and patched
onto SC-LEU-TRP-URA-HIS+3AT (40 mM) plates.
30

*Selecting inhibition of R4-FKBP interaction by FK506 using
the 5-FOA assay*

          The diploids isolated from the M x N analysis are
pooled and inoculated, into a medium that contains, instead
35 of glucose, a carbon source like lactate that does not induce
the expression from the ADH promoter (Denis et al., 1983, J.
Biol. Chem. 258:1165) that is driving transcription of the

two fusion genes. The medium also lacks tryptophan and
leucine, to maintain the GBD and GAD plasmids. This is
repeated in the presence or absence of FK506 at a final
concentration of 1 μM. This concentration of FK506 has been
5 demonstrated to inhibit the interaction of R4 with FKBP-12 in
the yeast two-hybrid system (Wang et al., 1994, Science
265:674-676).

The cells are grown in the Lactate medium (that
also lacks tryptophan and leucine) for 24-48 hours and then
10 diluted at a 1:100 ratio in SC-URA-LEU-TRP-HIS+FK506(1μM)
liquid media and grown for 8-24 hours. The carbon source in
this medium is glucose that supports the induction of
transcription from the ADH promoter (Holland and Holland,
1978, Biochemistry 17:4900). Growth is monitored by
15 measuring OD_{600}.

Dilutions of the culture are plated on
SC-LEU-TRP-HIS+FK506(1μM)+5-FOA plates and after a growth
period of 24-48 hours, fifty colonies from each dilution are
picked up and whole cell PCR is performed in parallel with
20 R4-specific primers (ALK5SfiI S (SEQ ID NO:45) and ALK5Sfi A
(SEQ ID NO:46)) and FKBP-12-specific primers (FKBPSfiI-A (SEQ
ID NO:48) and FKBPSfiI-S (SEQ ID NO:47). In this manner, the
selection of R4-FKBP diploids is monitored by the presence of
the specific PCR product. The ratio of R4-FKBP diploids to
25 the total number of diploids obtained indicates the degree of
enrichment of the FK506 inhibition of R4-FKBP interaction due
to 5-FOA selection.

The entire protocol is outlined in Figure 24.

30   6.18.   SELECTION OF NOVEL INTERACTING
                    PROTEINS AND INHIBITORS OF THESE
                    INTERACTING PROTEINS

The above example in Section 6.17 provides a means
to select for those yeast cells in which the interaction
between two proteins is inhibited by an inhibitor. A mixture
35 of cells that bear interacting proteins, that have risen from
an 'M x N' screen can be subjected to the above assay with

many inhibitors being screened against the mixture of cells bearing pairs of interacting proteins (Figure 6). Only those cells will survive in the 5-FOA media in which the inhibitors successfully inhibit the protein-protein interaction and
5 thereby do not activate the *URA3* reporter gene. This process can be iterative to enrich for a population of cells representing interacting pairs of proteins that are inhibited by specific inhibitors. The individual inhibition events can be sorted by diluting the cultures from 5-FOA media and
10 single colony purifying each diploid representing a pair of interacting proteins whose identity is confirmed by sequencing.

Isolation of interacting proteins from an M x N analysis
15   The M (YULH transformed with GBD fusions in pAS2-1) and N cells (N106' transformed with GAD fusions in pAD-GAL4) are mixed together and 1.0 ml of the mix (a total cell density of 1.5 x 10^8 cells/ml) is spread onto YPAD plates and incubated at 30°C for 6-8 hours for mating. A total of 1.7 x
20 10^9 cells representing 5 x 10^5 yeast transformants are present in the entire mating mix. The 5 x 10^5 yeast transformants arise from a library of 1 x 10^6 individual GBD or GAD fusion plasmids. These populations are sufficient to screen for interacting proteins form genes that are expressed at a level
25 of 1 in a 1000. The mating mixes are then harvested in 1 ml of SC-URA-LEU-TRP media and plated onto SC-URA-LEU- TRP-HIS plates that contain 40 mM 3-AT and incubated at 30°C until colonies emerge.
          Cells that are URA+, HIS+ and 3-AT resistant are
30 patched onto separate SC-TRP-LEU plates and assayed for β-galactosidase activity by the filter-lift assay. Cells that are URA+, HIS+, 3-AT resistant and positive for β-galactosidase activity are classified as positive for protein-protein interactions.
35

Selecting inhibitors of novel protein-protein interactions using the 5-FOA assay

- 291 -

Step I:

The diploids isolated from the M x N analysis are
pooled and inoculated into a medium that contains, instead
of glucose, a carbon source like lactate that does not induce
5 the expression from the ADH promoter (Denis et al., 1983, J.
Biol. Chem. 258:1165) that is driving transcription of the
two fusion genes. The medium also lacks tryptophan and
leucine to maintain the GBD and GAD plasmids. The cells are
inoculated in a a 96-well plate containing 150 μl of media.
10 Each well in the 96-well plate receives a unique inhibitor
compound. Each 96-well plate is repeated in triplicate with
each replicate receiving one particular concentration (1-100
μM) of the inhibitor compound.

15 Step II:

The cells are grown in the lactate medium for 24-48
hours at 30°C and then diluted at a 1:100 ratio in SC-URA-
LEU-TRP-HIS liquid media and grown for 8-24 hours. The carbon
source in this medium is glucose that supports the induction
20 of transcription from the ADH promoter (Holland and Holland,
1978, Biochemistry 17:4900). Growth is monitored by
measuring OD_{600}. As described above, the cells are inoculated
in a a 96-well plate containing 150 μl of media. Each well
in the 96-well plate receives a unique inhibitor compound.
25 Each 96-well plate is repeated in triplicate with each
replicate receiving one particular concentration (1-100 μM)
of the inhibitor compound.

Step III:

30   After this, a 1 to a 100 dilution of the cells is
transferred to similar 96-well plates that contain
SC-LEU-TRP+5-FOA liquid media (150 μl). The chemicals
(identity and concentration) present in each well are
identical to that present in Step II. The cells are
35 incubated at 30°C for 8-48 hours.

Step IV:

After proper mixing, 5 µl of cells from each well
is spotted onto a plate with the same dimensions as the
96-well plate and containing SC-URA-LEU-TRP-HIS agar and
5 incubated at 30°C. After 2-4 days, colonies of yeast grow up
and these are picked and patched onto SC-LEU-TRP media.

Step V:

A confirmation of the inhibition of the pooled
10 interactants by the particular inhibitor is performed by
inoculating, in parallel, the colonies from each patch in
Step IV to a well containing SC-URA-LEU-TRP-HIS liquid media
and SC-URA-LEU-TRP-HIS liquid media with the same inhibitor
at an identical concentration (i.e., as in Step III) that
15 gave rise to 5-FOA resistant colonies. The cultures are
incubated at 30°C for 24-48 hours, and growth is monitored by
measuring OD_{600}. Inhibition of growth should be observed in
the presence of the inhibitor, while none should be evident
in the absence of the inhibitor. β-galactosidase activity is
20 measured in a fraction of the cells using the FluoReporter,
lacZ/Galactosidase Quantitation kit (Molecular Probes)
according to the manufacturers protocols and a decrease in
β-galactosidase activity should be observed in the presence
of the inhibitor in comparison to the cells grown in the
25 absence of the inhibitor.

*Identification of the pairs of interacting proteins that are
inhibited by specific inhibitors*
Whole cell PCR is performed, as described in
30 Section 6.1.8, on the colonies that are isolated as a result
of the 5-FOA selection. This is done in parallel with both
the GBD-fusion plasmid specific and GAD-fusion plasmid
specific primer pairs. If more than one PCR product is
observed from one patch of cells, it indicates that more than
35 one pair of interacting proteins are inhibited by the same
inhibitor. Then, the patch of colonies are streak-purified to
yield clonal colonies and the whole cell PCR procedure is

- 293 -

repeated. The presence of a single PCR product confirms the clonal nature of the colony. The PCR products are identified to reveal the identity of the genes encoding the pair of interacting proteins.

5   Thus the above method (outlined in Figure 25) provides a high throughput mechanism for isolation of inhibitors against all possible pairs of interacting proteins that are characteristic to a particular population, be it a cell-type, disease-state or stage of development.

10

          The present invention is not to be limited in scope by the specific embodiments described herein. Indeed, various modifications of the invention in addition to those described herein will become apparent to those skilled in the

15 art from the foregoing description and accompanying figures. Such modifications are intended to fall within the scope of the appended claims.

          Various publications are cited herein, the disclosures of which are incorporated by reference in their

20 entireties.

25

30

35

SEQUENCE LISTING

(1) GENERAL INFORMATION

(i) APPLICANT: Nandabalan, Krishnan
Rothberg, Jonathan
Yang, Meijia
Knight, James R.
Kalbfleisch, Theodore S.

(ii) TITLE OF THE INVENTION: IDENTIFICATION AND COMPARISON OF
PROTEIN-PROTEIN INTERACTIONS THAT OCCUR IN POPULATIONS
AND IDENTIFICATION OF INHIBITORS OF THESE INTERACTORS

(iii) NUMBER OF SEQUENCES: 122

(iv) CORRESPONDENCE ADDRESS:
    (A) ADDRESSEE: Pennie & Edmonds
    (B) STREET: 1155 Avenue of the Americas
    (C) CITY: New York
    (D) STATE: NY
    (E) COUNTRY: USA
    (F) ZIP: 10036/2711

(v) COMPUTER READABLE FORM:
    (A) MEDIUM TYPE: Diskette
    (B) COMPUTER: IBM Compatible
    (C) OPERATING SYSTEM: DOS
    (D) SOFTWARE: FastSEQ Version 2.0

(vi) CURRENT APPLICATION DATA:
    (A) APPLICATION NUMBER:
    (B) FILING DATE: 13-JUN-1997
    (C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:
    (A) APPLICATION NUMBER: 08/663,824
    (B) FILING DATE: 14-JUN-1996

(viii) ATTORNEY/AGENT INFORMATION:
    (A) NAME: Misrock, S. Leslie
    (B) REGISTRATION NUMBER: 18,872
    (C) REFERENCE/DOCKET NUMBER: 7934-045

(ix) TELECOMMUNICATION INFORMATION:
    (A) TELEPHONE: 212-790-9090
    (B) TELEFAX: 212-869-8864
    (C) TELEX: 66141 PENNIE

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 21 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

CGATGCACAG TTGAAGTGAA C   21

          (2) INFORMATION FOR SEQ ID NO:2:

          (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 25 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear
5
          (ii) MOLECULE TYPE: DNA

          (xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

     CGCGTTTGGA ATCACTACAG GGATG   25

              (2) INFORMATION FOR SEQ ID NO:3:

10   (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 36 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

          (ii) MOLECULE TYPE: DNA

          (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:
15
     CTACCAGAAT TCGGCATGCC GGTAGAGGTG TGGTCA   36

              (2) INFORMATION FOR SEQ ID NC:4:

          (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 25 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
20   (D) TOPOLOGY: linear

          (ii) MOLECULE TYPE: DNA

          (xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

     ATGAAGCTAC TGTCTTCTAT CGAAC   25

              (2) INFORMATION FOR SEQ ID NO:5:
25
          (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 48 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

          (ii) MOLECULE TYPE: DNA

30   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

     CATATGGCCG AGGTGGCCTA GGGCCTCCTG GGCCTCCCTT AGGGATCC   48

              (2) INFORMATION FOR SEQ ID NO:6:

          (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 42 base pairs
            (B) TYPE: nucleic acid
35   (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

          (ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

GAGGCCGAGG TGGCCTAGGG CCTCCTGGGC CTCTAGAATT CC   42

    (2) INFORMATION FOR SEQ ID NO:7:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 43 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

GGACTAGGCC GAGGTGGCCG GTATGACGGA ATATAAGCTG GTG   43

    (2) INFORMATION FOR SEQ ID NO:8:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 40 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

GGACTAGGCC GAGGTGGCCG GAGAGCACAC ACTTGCAGCT   40

    (2) INFORMATION FOR SEQ ID NO:9:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 40 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

GGACTAGGCC GAGGTGGCCA TGGAGCACAT ACAGGGAGCT   40

    (2) INFORMATION FOR SEQ ID NO:10:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 40 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

GGACTAGGCC GAGGTGGCCC GACCTCTGCC TCTGGGAGAG   40

    (2) INFORMATION FOR SEQ ID NO:11:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 40 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single

                    (D) TOPOLOGY: linear

                (ii) MOLECULE TYPE: DNA

                (xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

         GGACTAGGCC GAGGTGGCCG GAGGAGGGCA GAATCATCAC   40

 5   (2) INFORMATION FOR SEQ ID NO:12:

                (i) SEQUENCE CHARACTERISTICS:
                    (A) LENGTH: 42 base pairs
                    (B) TYPE: nucleic acid
                    (C) STRANDEDNESS: single
                    (D) TOPOLOGY: linear

10   (ii) MOLECULE TYPE: DNA

                (xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

         GGACTAGGCC TCCTGGGCCA CGCCTCGGCT TGTCACATCT GC   42

                    (2) INFORMATION FOR SEQ ID NO:13:

                (i) SEQUENCE CHARACTERISTICS:
15   (A) LENGTH: 45 base pairs
                    (B) TYPE: nucleic acid
                    (C) STRANDEDNESS: single
                    (D) TOPOLOGY: linear

                (ii) MOLECULE TYPE: DNA

                (xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

20 GGACTAGGCC GAGGTGGCCC TCTCTGTGGG TTTGCCTAGT GTTTC   45

                    (2) INFORMATION FOR SEQ ID NO:14:

                (i) SEQUENCE CHARACTERISTICS:
                    (A) LENGTH: 43 base pairs
                    (B) TYPE: nucleic acid
                    (C) STRANDEDNESS: single
                    (D) TOPOLOGY: linear
25
                (ii) MOLECULE TYPE: DNA

                (xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

         GGACTAGGCC TCCTGGGCCC TCCTTTGAAA TGGGATTGGT AAG   43

                    (2) INFORMATION FOR SEQ ID NO:15:

30   (i) SEQUENCE CHARACTERISTICS:
                    (A) LENGTH: 11 base pairs
                    (B) TYPE: nucleic acid
                    (C) STRANDEDNESS: single
                    (D) TOPOLOGY: linear

                (ii) MOLECULE TYPE: DNA

                (xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:
35
         AGGCCGGAGG C   11

                    (2) INFORMATION FOR SEQ ID NO:16:

        (i) SEQUENCE CHARACTERISTICS:
         (A) LENGTH: 14 base pairs
         (B) TYPE: nucleic acid
         (C) STRANDEDNESS: single
         (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

5
        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

TCCTCCGGCC TCCG   14

            (2) INFORMATION FOR SEQ ID NO:17:

        (i) SEQUENCE CHARACTERISTICS:
         (A) LENGTH: 27 base pairs
10   (B) TYPE: nucleic acid
         (C) STRANDEDNESS: single
         (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

ACGTGCAAGG CCCAGGAGGC CGGAGGC   27
15
            (2) INFORMATION FOR SEQ ID NO:18:

        (i) SEQUENCE CHARACTERISTICS:
         (A) LENGTH: 38 base pairs
         (B) TYPE: nucleic acid
         (C) STRANDEDNESS: single
         (D) TOPOLOGY: linear

20   (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

GGGACAAACG GCCGCACCGA AACGCGCGAG GCAGCAAC   38

            (2) INFORMATION FOR SEQ ID NO:19:

        (i) SEQUENCE CHARACTERISTICS:
25   (A) LENGTH: 37 base pairs
         (B) TYPE: nucleic acid
         (C) STRANDEDNESS: single
         (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

30 GGGAGTTGCA TGCGCCGGTA GAGGTGTGGT CAATAAG   37

            (2) INFORMATION FOR SEQ ID NO:20:

        (i) SEQUENCE CHARACTERISTICS:
         (A) LENGTH: 7 amino acids
         (B) TYPE: amino acid
         (C) STRANDEDNESS: unknown
         (D) TOPOLOGY: unknown
35
        (ii) MOLECULE TYPE: peptide

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

Pro Lys Lys Lys Arg Lys Val
1   5

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 56 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

AGCTTGGCCT CCCAGGCCAC AGACAGGCGC GCCCCCAAAG AAGAGAAAGG TTTAGA   56

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 59 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

AGCTTCTAAA CCTTTCTCTT CTTCTTTGGG GGCGCGCCTG TCTGTGGCCT GGGAGGCCA   59

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 43 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

GGACTAGGCC GAGGTGGCCG GTATGACGGA ATATAAGCTG GTG   43

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 40 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

GGACTAGGCC GAGGTGGCCG GAGAGCACAC ACTTGCAGCT   40

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 40 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

GGACTAGGCC GAGGTGGCCA TGGAGCACAT ACAGGGAGCT   40

     (2) INFORMATION FOR SEQ ID NO:26:

5

(i) SEQUENCE CHARACTERISTICS:
   (A) LENGTH: 40 base pairs
   (B) TYPE: nucleic acid
   (C) STRANDEDNESS: single
   (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

10

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

GGACTAGGCC GAGGTGGCCC GACCTCTGCC TCTGGGAGAG   40

     (2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS:
   (A) LENGTH: 24 base pairs
   (B) TYPE: nucleic acid
15   (C) STRANDEDNESS: single
   (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

AGCACTCTCC AGCCTCTCAC CGAC   24

20   (2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:
   (A) LENGTH: 12 base pairs
   (B) TYPE: nucleic acid
   (C) STRANDEDNESS: single
   (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

25

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

GATCGTCGGT GA   12

     (2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS:
   (A) LENGTH: 24 base pairs
30   (B) TYPE: nucleic acid
   (C) STRANDEDNESS: single
   (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

AGCACTCTCC AGCCTCTCAC CGAC   24

35

     (2) INFORMATION FOR SEQ ID NO:30:

(i) SEQUENCE CHARACTERISTICS:
   (A) LENGTH: 12 base pairs

            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

CCGGGTCGGT GA   12

        (2) INFORMATION FOR SEQ ID NO:31:

        (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 24 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

AGCACTCTCC AGCCTCTCAC CGAC   24

        (2) INFORMATION FOR SEQ ID NO:32:

        (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 12 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:

TTAAGTCGGT GA   12

        (2) INFORMATION FOR SEQ ID NO:33:

        (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 24 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:

AGCACTCTCC AGCCTCTCAC CGAC   24

        (2) INFORMATION FOR SEQ ID NO:34:

        (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 12 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:

AGCTGTCGGT GA   12

(2) INFORMATION FOR SEQ ID NO:35:

   (i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 21 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

5

   (ii) MOLECULE TYPE: RNA

   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:

CCCUGAUGGU AGACCGGGGU G   21

(2) INFORMATION FOR SEQ ID NO:36:

10
   (i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 50 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

   (ii) MOLECULE TYPE: DNA

   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:

15
GAGAGAGAGA GAGAGAGAGA ACTAGTCTCG AGTTTTTTTT TTTTTTTTTT   50

(2) INFORMATION FOR SEQ ID NO:37:

   (i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 13 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
20
    (D) TOPOLOGY: linear

   (ii) MOLECULE TYPE: DNA

   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:

AATTCGGCAC GAG   13

(2) INFORMATION FOR SEQ ID NO:38:

25
   (i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 9 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

   (ii) MOLECULE TYPE: DNA

30
   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:38:

CTCGTGCCG   9

(2) INFORMATION FOR SEQ ID NO:39:

   (i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 47 base pairs
    (B) TYPE: nucleic acid
35
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

   (ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:

GAGAGAGAGA GGGTACCGAA CCAATGTATC CAGCACCACC TGTAACC   47

      (2) INFORMATION FOR SEQ ID NO:40:

    (i) SEQUENCE CHARACTERISTICS:
     (A) LENGTH: 52 base pairs
     (B) TYPE: nucleic acid
     (C) STRANDEDNESS: single
     (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:

GAGACAGAGA ATTCCATTAT AGTTTTTTCT CCTTGACGTT AAAGTATAGA GG   52

      (2) INFORMATION FOR SEQ ID NO:41:

    (i) SEQUENCE CHARACTERISTICS:
     (A) LENGTH: 43 base pairs
     (B) TYPE: nucleic acid
     (C) STRANDEDNESS: single
     (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:41:

GAGAGACAGA ATTCTCGAAA GCTACATATA AGGAACGTGC TGC   43

      (2) INFORMATION FOR SEQ ID NO:42:

    (i) SEQUENCE CHARACTERISTICS:
     (A) LENGTH: 42 base pairs
     (B) TYPE: nucleic acid
     (C) STRANDEDNESS: single
     (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:42:

GAGAGAGACG GCCGCGTCAT TATAGAAATC ATTACGACCG AG   42

      (2) INFORMATION FOR SEQ ID NO:43:

    (i) SEQUENCE CHARACTERISTICS:
     (A) LENGTH: 90 base pairs
     (B) TYPE: nucleic acid
     (C) STRANDEDNESS: single
     (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:43:

GATTCGGTAA TCTCCGAACA GAAGGAAGAA CGAAGGAAGG AGCACAGACT TAGATTGGTA   60
GAACCAATGT ATCCAGCACC ACCTGTAACC   90

      (2) INFORMATION FOR SEQ ID NO:44:

    (i) SEQUENCE CHARACTERISTICS:
     (A) LENGTH: 39 base pairs
     (B) TYPE: nucleic acid

        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:44:

5   ACATCAAAAG GCCTCTAGGT TCCTTTGTTA CTTCTTCCG   39

        (2) INFORMATION FOR SEQ ID NO:45:

        (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 43 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear
10
    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:45:

    GGACTAGGCC GAGGTGGCCT GCCACAACCG CACTGTCATT CAC   43

        (2) INFORMATION FOR SEQ ID NO:46:

15   (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 49 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:46:
20
    GGACTAGGCC TCCTGGGCCT TACATTTTGA TGCCTTCCTG TTGACTGAG   49

        (2) INFORMATION FOR SEQ ID NO:47:

        (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 43 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
25   (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:47:

    GGACTAGGCC GAGGTGGCCA TGGGAGTGCA GGTGGAAACC ATC   43

        (2) INFORMATION FOR SEQ ID NO:48:
30
        (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 43 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

35 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:48:

    GGACTAGGCC TCCTGGGCCT CATTCCAGTT TTAGAAGCTC CAC   43

        (2) INFORMATION FOR SEQ ID NO:49:

```
              (i) SEQUENCE CHARACTERISTICS:
                 (A) LENGTH: 20 base pairs
                 (B) TYPE: nucleic acid
                 (C) STRANDEDNESS: single
                 (D) TOPOLOGY: linear

             (ii) MOLECULE TYPE: DNA
  5
             (xi) SEQUENCE DESCRIPTION: SEQ ID NO:49:

      TTGGAATCAC TACAGGGATG                              20

              (2) INFORMATION FOR SEQ ID NO:50:

              (i) SEQUENCE CHARACTERISTICS:
                 (A) LENGTH: 21 base pairs
 10              (B) TYPE: nucleic acid
                 (C) STRANDEDNESS: single
                 (D) TOPOLOGY: linear

             (ii) MOLECULE TYPE: DNA

             (xi) SEQUENCE DESCRIPTION: SEQ ID NO:50:

      GAATTCATGG CTTACCCATA C                            21
 15
              (2) INFORMATION FOR SEQ ID NO:51:

              (i) SEQUENCE CHARACTERISTICS:
                 (A) LENGTH: 25 base pairs
                 (B) TYPE: nucleic acid
                 (C) STRANDEDNESS: single
                 (D) TOPOLOGY: linear

 20          (ii) MOLECULE TYPE: DNA

             (xi) SEQUENCE DESCRIPTION: SEQ ID NO:51:

      AACCTGACCT ACAGGAAAGA GTTAC                        25

              (2) INFORMATION FOR SEQ ID NO:52:

              (i) SEQUENCE CHARACTERISTICS:
 25              (A) LENGTH: 23 base pairs
                 (B) TYPE: nucleic acid
                 (C) STRANDEDNESS: single
                 (D) TOPOLOGY: linear

             (ii) MOLECULE TYPE: DNA

             (xi) SEQUENCE DESCRIPTION: SEQ ID NO:52:

 30   CCTCTAACAT TGAGACAGCA TAG                          23

              (2) INFORMATION FOR SEQ ID NO:53:

              (i) SEQUENCE CHARACTERISTICS:
                 (A) LENGTH: 24 base pairs
                 (B) TYPE: nucleic acid
                 (C) STRANDEDNESS: single
                 (D) TOPOLOGY: linear
 35
             (ii) MOLECULE TYPE: DNA

             (xi) SEQUENCE DESCRIPTION: SEQ ID NO:53:
```

AGCACTCTCC AGCCTCTCAC CGAA   24

        (2) INFORMATION FOR SEQ ID NO:54:

      (i) SEQUENCE CHARACTERISTICS:
       (A) LENGTH: 12 base pairs
       (B) TYPE: nucleic acid
5   (C) STRANDEDNESS: single
       (D) TOPOLOGY: linear

      (ii) MOLECULE TYPE: DNA

      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:54:

AATTTTCGGT GA   12

10   (2) INFORMATION FOR SEQ ID NO:55:

      (i) SEQUENCE CHARACTERISTICS:
       (A) LENGTH: 12 base pairs
       (B) TYPE: nucleic acid
       (C) STRANDEDNESS: single
       (D) TOPOLOGY: linear

      (ii) MOLECULE TYPE: DNA
15
      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:55:

CATGTTCGGT GA   12

        (2) INFORMATION FOR SEQ ID NO:56:

      (i) SEQUENCE CHARACTERISTICS:
       (A) LENGTH: 12 base pairs
20   (B) TYPE: nucleic acid
       (C) STRANDEDNESS: single
       (D) TOPOLOGY: linear

      (ii) MOLECULE TYPE: DNA

      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:56:

CCGGTTCGGT GA   12
25
       (2) INFORMATION FOR SEQ ID NO:57:

      (i) SEQUENCE CHARACTERISTICS:
       (A) LENGTH: 12 base pairs
       (B) TYPE: nucleic acid
       (C) STRANDEDNESS: single
       (D) TOPOLOGY: linear

30   (ii) MOLECULE TYPE: DNA

      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:57:

CGCGTTCGGT GA   12

       (2) INFORMATION FOR SEQ ID NO:58:

      (i) SEQUENCE CHARACTERISTICS:
35   (A) LENGTH: 12 base pairs
       (B) TYPE: nucleic acid
       (C) STRANDEDNESS: single
       (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:58:

CTAGTTCGGT GA   12

    (2) INFORMATION FOR SEQ ID NO:59:

5

(i) SEQUENCE CHARACTERISTICS:
  (A) LENGTH: 12 base pairs
  (B) TYPE: nucleic acid
  (C) STRANDEDNESS: single
  (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

10   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:59:

GATCTTCCGT GA   12

    (2) INFORMATION FOR SEQ ID NO:60:

(i) SEQUENCE CHARACTERISTICS:
  (A) LENGTH: 12 base pairs
  (B) TYPE: nucleic acid
15 (C) STRANDEDNESS: single
  (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCPIPTION: SEQ ID NO:60:

GCGCTTCCGT GA   12

20   (2) INFORMATION FOR SEQ ID NO:61:

(i) SEQUENCE CHARACTERISTICS:
  (A) LENGTH: 12 base pairs
  (B) TYPE: nucleic acid
  (C) STRANDEDNESS: single
  (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

25

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:61:

GGCCTTCGGT GA   12

    (2) INFORMATION FOR SEQ ID NO:62:

(i) SEQUENCE CHARACTERISTICS:
  (A) LENGTH: 12 base pairs
30 (B) TYPE: nucleic acid
  (C) STRANDEDNESS: single
  (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:62:

GTACTTCGGT GA   12
35

    (2) INFORMATION FOR SEQ ID NO:63:

(i) SEQUENCE CHARACTERISTICS:
  (A) LENGTH: 12 base pairs

                    (B) TYPE: nucleic acid
                    (C) STRANDEDNESS: single
                    (D) TOPOLOGY: linear

                (ii) MOLECULE TYPE: DNA

                (xi) SEQUENCE DESCRIPTION: SEQ ID NO:63:
 5
        TCGATTCGGT GA   12

                (2) INFORMATION FOR SEQ ID NO:64:

                (i) SEQUENCE CHARACTERISTICS:
                    (A) LENGTH: 12 base pairs
                    (B) TYPE: nucleic acid
                    (C) STRANDEDNESS: single
10   (D) TOPOLOGY: linear

                (ii) MOLECULE TYPE: DNA

                (xi) SEQUENCE DESCRIPTION: SEQ ID NO:64:

        TGCATTCGGT GA   12

                (2) INFORMATION FOR SEQ ID NO:65:
15
                (i) SEQUENCE CHARACTERISTICS:
                    (A) LENGTH: 12 base pairs
                    (B) TYPE: nucleic acid
                    (C) STRANDEDNESS: single
                    (D) TOPOLOGY: linear

                (ii) MOLECULE TYPE: DNA

20   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:65:

        TTAATTCGGT GA   12

                (2) INFORMATION FOR SEQ ID NO:66:

                (i) SEQUENCE CHARACTERISTICS:
                    (A) LENGTH: 12 base pairs
                    (B) TYPE: nucleic acid
25   (C) STRANDEDNESS: single
                    (D) TOPOLOGY: linear

                (ii) MOLECULE TYPE: DNA

                (xi) SEQUENCE DESCRIPTION: SEQ ID NO:66:

        ACGATTCGGT GA   12

30   (2) INFORMATION FOR SEQ ID NO:67:

                (i) SEQUENCE CHARACTERISTICS:
                    (A) LENGTH: 24 base pairs
                    (B) TYPE: nucleic acid
                    (C) STRANDEDNESS: single
                    (D) TOPOLOGY: linear

                (ii) MOLECULE TYPE: DNA
35
                (xi) SEQUENCE DESCRIPTION: SEQ ID NO:67:

        AGCACTCTCC AGCCTCTCAC CGAC   24

(2) INFORMATION FOR SEQ ID NO:68:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

5

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:68:

AATTGTCGCT GA   12

(2) INFORMATION FOR SEQ ID NO:69:

10   (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:69:

15
AGCTGTCGCT GA   12

(2) INFORMATION FOR SEQ ID NO:70:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
20   (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:70:

CATGGTCGCT GA   12

(2) INFORMATION FOR SEQ ID NO:71:

25
    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

30   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:71:

CCGGGTCGCT GA   12

(2) INFORMATION FOR SEQ ID NO:72:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
35   (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:72:

CGCGGTCGCT GA   12

     (2) INFORMATION FOR SEQ ID NO:73:

    (i) SEQUENCE CHARACTERISTICS:
     (A) LENGTH: 12 base pairs
     (B) TYPE: nucleic acid
     (C) STRANDEDNESS: single
     (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:73:

CTAGGTCGCT GA   12

     (2) INFORMATION FOR SEQ ID NO:74:

    (i) SEQUENCE CHARACTERISTICS:
     (A) LENGTH: 12 base pairs
     (B) TYPE: nucleic acid
     (C) STRANDEDNESS: single
     (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:74:

GATCGTCGCT GA   12

     (2) INFORMATION FOR SEQ ID NO:75:

    (i) SEQUENCE CHARACTERISTICS:
     (A) LENGTH: 12 base pairs
     (B) TYPE: nucleic acid
     (C) STRANDEDNESS: single
     (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:75:

GCGCGTCGCT GA   12

     (2) INFORMATION FOR SEQ ID NO:76:

    (i) SEQUENCE CHARACTERISTICS:
     (A) LENGTH: 12 base pairs
     (B) TYPE: nucleic acid
     (C) STRANDEDNESS: single
     (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:76:

GGCCGTCGCT GA   12

     (2) INFORMATION FOR SEQ ID NO:77:

    (i) SEQUENCE CHARACTERISTICS:
     (A) LENGTH: 12 base pairs
     (B) TYPE: nucleic acid
     (C) STRANDEDNESS: single

        (D) TOPOLOGY: linear

     (ii) MOLECULE TYPE: DNA

     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:77:

GTACGTCGCT GA   12

5
        (2) INFORMATION FOR SEQ ID NO:78:

     (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

10   (ii) MOLECULE TYPE: DNA

     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:78:

TCGAGTCGCT GA   12

        (2) INFORMATION FOR SEQ ID NO:79:

     (i) SEQUENCE CHARACTERISTICS:
15   (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

     (ii) MOLECULE TYPE: DNA

     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:79:

20 TGCAGTCGCT GA   12

        (2) INFORMATION FOR SEQ ID NO:80:

     (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 24 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

25
     (ii) MOLECULE TYPE: DNA

     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:80:

ACCGACGTCG ACTATCCATG AAGA   24

        (2) INFORMATION FOR SEQ ID NO:81:

30   (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

     (ii) MOLECULE TYPE: DNA

     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:81:
35
AATTTCTTCA TG   12

        (2) INFORMATION FOR SEQ ID NO:82:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 12 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

5

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:82:

CATGTCTTCA TG   12

        (2) INFORMATION FOR SEQ ID NO:83:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 12 base pairs
10   (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:83:

CCGGTCTTCA TG   12

15
        (2) INFORMATION FOR SEQ ID NO:84:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 12 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

20   (ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:84:

CGCGTCTTCA TG   12

        (2) INFORMATION FOR SEQ ID NO:85:

(i) SEQUENCE CHARACTERISTICS:
25   (A) LENGTH: 12 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:85:

30 CTAGTCTTCA TG   12

        (2) INFORMATION FOR SEQ ID NO:86:

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 12 base pairs
    (B) TYPE: nucleic acid
    (C) STRANDEDNESS: single
    (D) TOPOLOGY: linear

35
(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:86:

```
GATCTCTTCA TG                                                        12

        (2) INFORMATION FOR SEQ ID NO:87:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
5     (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:87:

GCGCTCTTCA TG                                                        12

10      (2) INFORMATION FOR SEQ ID NO:88:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA
15
    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:88:

GGCCTCTTCA TG                                                        12

        (2) INFORMATION FOR SEQ ID NO:89:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
20    (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:89:

GTACTCTTCA TG                                                        12
25
        (2) INFORMATION FOR SEQ ID NO:90:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

30  (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:90:

TCGATCTTCA TG                                                        12

        (2) INFORMATION FOR SEQ ID NO:91:

    (i) SEQUENCE CHARACTERISTICS:
35    (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear
```

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:91:

TGCATCTTCA TG   12

      (2) INFORMATION FOR SEQ ID NO:92:

5

(i) SEQUENCE CHARACTERISTICS:
  (A) LENGTH: 12 base pairs
  (B) TYPE: nucleic acid
  (C) STRANDEDNESS: single
  (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

10   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:92:

TTAATCTTCA TG   12

      (2) INFORMATION FOR SEQ ID NO:93:

(i) SEQUENCE CHARACTERISTICS:
  (A) LENGTH: 12 base pairs
  (B) TYPE: nucleic acid
15 (C) STRANDEDNESS: single
  (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:93:

ACGATCTTCA TG   12

20   (2) INFORMATION FOR SEQ ID NO:94:

(i) SEQUENCE CHARACTERISTICS:
  (A) LENGTH: 24 base pairs
  (B) TYPE: nucleic acid
  (C) STRANDEDNESS: single
  (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

25
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:94:

ACCGACGTCG ACTATCCATG AAGC   24

      (2) INFORMATION FOR SEQ ID NO:95:

(i) SEQUENCE CHARACTERISTICS:
  (A) LENGTH: 12 base pairs
30 (B) TYPE: nucleic acid
  (C) STRANDEDNESS: single
  (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:95:

AATTGCTTCA TG   12

35
      (2) INFORMATION FOR SEQ ID NO:96:

(i) SEQUENCE CHARACTERISTICS:
  (A) LENGTH: 12 base pairs

```
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

       (ii) MOLECULE TYPE: DNA

       (xi) SEQUENCE DESCRIPTION: SEQ ID NO:96:
  5
     AGCTGCTTCA TG                                              12

             (2) INFORMATION FOR SEQ ID NO:97:

          (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 12 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
 10       (D) TOPOLOGY: linear

       (ii) MOLECULE TYPE: DNA

       (xi) SEQUENCE DESCRIPTION: SEQ ID NO:97:

     CATGGCTTCA TG                                              12

             (2) INFORMATION FOR SEQ ID NO:98:
 15
          (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 12 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

       (ii) MOLECULE TYPE: DNA

 20       (xi) SEQUENCE DESCRIPTION: SEQ ID NO:98:

     CCGGGCTTCA TG                                              12

             (2) INFORMATION FOR SEQ ID NO:99:

          (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 12 base pairs
          (B) TYPE: nucleic acid
 25       (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

       (ii) MOLECULE TYPE: DNA

       (xi) SEQUENCE DESCRIPTION: SEQ ID NO:99:

     CGCGGCTTCA TG                                              12

 30          (2) INFORMATION FOR SEQ ID NO:100:

          (i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 12 base pairs
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

       (ii) MOLECULE TYPE: DNA
 35
       (xi) SEQUENCE DESCRIPTION: SEQ ID NO:100:

     CATGGCTTCA TG                                              12
```

(2) INFORMATION FOR SEQ ID NO:101:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

5

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:101:

GATCGCTTCA TG   12

    (2) INFORMATION FOR SEQ ID NO:102:

10
    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:102:

15
GCGCGCTTCA TG   12

    (2) INFORMATION FOR SEQ ID NO:103:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
20
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:103:

GGCCGCTTCA TG   12

    (2) INFORMATION FOR SEQ ID NO:104:

25
    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

30   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:104:

GTACGCTTCA TG   12

    (2) INFORMATION FOR SEQ ID NO:105:

    (i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 12 base pairs
      (B) TYPE: nucleic acid
35   (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

    (ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:105:

TCGAGCTTCA TG   12

        (2) INFORMATION FOR SEQ ID NO:106:

        (i) SEQUENCE CHARACTERISTICS:
5   (A) LENGTH: 12 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:106:

10 TGCAGCTTCA TG   12

        (2) INFORMATION FOR SEQ ID NO:107:

        (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 24 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear
15
        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NC:107:

AGCACTCTCC AGCCUCTCAC CGAA   24

        (2) INFORMATION FOR SEQ ID NO:108:

20   (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 24 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
            (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:108:
25
AGCACTCTGG CGCGCCTCAC CGAA   24

        (2) INFORMATION FOR SEQ ID NO:109:

        (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 24 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single
30   (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:109:

AGCACTCTCC AGCCUCTCAC CGAC   24

        (2) INFORMATION FOR SEQ ID NO:110:
35
        (i) SEQUENCE CHARACTERISTICS:
            (A) LENGTH: 24 base pairs
            (B) TYPE: nucleic acid
            (C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:110:

AGCACTCTGG CGCGCCTCAC CGAC   24

5
        (2) INFORMATION FOR SEQ ID NO:111:

        (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 24 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

10   (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:111:

ACCGACGTCG ACTATGGATG AAGA   24

        (2) INFORMATION FOR SEQ ID NO:112:

        (i) SEQUENCE CHARACTERISTICS:
15   (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:112:

20 GATCTCTTCA TC   12

        (2) INFORMATION FOR SEQ ID NO:113:

        (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

25
        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:113:

CATGTCTTCA TC   12

        (2) INFORMATION FOR SEQ ID NO:114:

30   (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 12 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

        (ii) MOLECULE TYPE: DNA

        (xi) SEQUENCE DESCRIPTION: SEQ ID NO:114:
35
CCGGTCTTCA TC   12

        (2) INFORMATION FOR SEQ ID NO:115:

         (i) SEQUENCE CHARACTERISTICS:
           (A) LENGTH: 21 base pairs
           (B) TYPE: nucleic acid
           (C) STRANDEDNESS: single
           (D) TOPOLOGY: linear

         (ii) MOLECULE TYPE: DNA

 5   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:115:

     ACCGACGTCG ACTATCGCAG C   21

           (2) INFORMATION FOR SEQ ID NO:116:

         (i) SEQUENCE CHARACTERISTICS:
           (A) LENGTH: 12 base pairs
10   (B) TYPE: nucleic acid
           (C) STRANDEDNESS: single
           (D) TOPOLOGY: linear

         (ii) MOLECULE TYPE: DNA

         (xi) SEQUENCE DESCRIPTION: SEQ ID NO:116:

     GATCTCTGCT GC   12
15
           (2) INFORMATION FOR SEQ ID NO:117:

         (i) SEQUENCE CHARACTERISTICS:
           (A) LENGTH: 12 base pairs
           (B) TYPE: nucleic acid
           (C) STRANDEDNESS: single
           (D) TOPOLOGY: linear

20   (ii) MOLECULE TYPE: DNA

         (xi) SEQUENCE DESCRIPTION: SEQ ID NO:117:

     CATGTCTGCT GC   12

           (2) INFORMATION FOR SEQ ID NO:118:

         (i) SEQUENCE CHARACTERISTICS:
25   (A) LENGTH: 39 base pairs
           (B) TYPE: nucleic acid
           (C) STRANDEDNESS: single
           (D) TOPOLOGY: linear

         (ii) MOLECULE TYPE: DNA

         (xi) SEQUENCE DESCRIPTION: SEQ ID NO:118:

30   AGGAAACAGC TATGACCATC TGAGAAAGCA ACCTGACCT   39

           (2) INFORMATION FOR SEQ ID NO:119:

         (i) SEQUENCE CHARACTERISTICS:
           (A) LENGTH: 39 base pairs
           (B) TYPE: nucleic acid
           (C) STRANDEDNESS: single
           (D) TOPOLOGY: linear
35
         (ii) MOLECULE TYPE: DNA

         (xi) SEQUENCE DESCRIPTION: SEQ ID NO:119:

GTTTTCCCAG TCACGACGGT GCGACATCAT CATCGGAAG   39

      (2) INFORMATION FOR SEQ ID NO:120:

      (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 39 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

      (ii) MOLECULE TYPE: DNA

      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:120:

GTTTTCCCAG TCACGACGAG GGATGTTTAA TACCACTAC   39

      (2) INFORMATION FOR SEQ ID NO:121:

      (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 40 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

      (ii) MOLECULE TYPE: DNA

      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:121:

AGGAAACAGC TATGACCATG CACAGTTGAA GTGAACTTGC   40

      (2) INFORMATION FOR SEQ ID NO:122:

      (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 27 base pairs
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

      (ii) MOLECULE TYPE: DNA

      (xi) SEQUENCE DESCRIPTION: SEQ ID NO:122:

ATGGATGATG TATATAACTA TCTATTC   27

WHAT IS CLAIMED IS:

1. A method of detecting one or more protein-protein interactions comprising

5   (a) recombinantly expressing within a population of host cells

         (i) a first population of first fusion proteins, each said first fusion protein comprising a first protein sequence and a DNA binding domain in which the

10 DNA binding domain is the same in each said first fusion protein, and in which said first population of first fusion proteins has a complexity of at least 1,000; and

         (ii) a second population of second fusion proteins, each said second fusion protein comprising a second

15 protein sequence and a transcriptional regulatory domain of a transcriptional regulator, in which the transcriptional regulatory domain is the same in each said second fusion protein, such that a first fusion protein is co-expressed with a second fusion protein in

20 host cells, and wherein said host cells contain at least one nucleotide sequence operably linked to a promoter driven by one or more DNA binding sites recognized by said DNA binding domain such that interaction of a first fusion protein with a second fusion protein results in

25 regulation of transcription of said at least one nucleotide sequence by said regulatory domain, and in which said second population of second fusion proteins has a complexity of at least 1,000; and

         (b) detecting said regulation of transcription of

30 said at least one nucleotide sequence, thereby detecting an interaction between a first fusion protein and a second fusion protein.

2. The method according to claim 1 in which the

35 regulatory domain is an activation domain, and said regulation of transcription is activation of transcription.

3. The method according to claim 1 in which the first
and second populations of fusion proteins are each expressed
from chimeric genes comprising cDNA sequences from an
uncharacterized sample of a population of cDNA from mammalian
5 RNA.

4. The method according to claim 2 in which the first
and second populations of fusion proteins comprise first and
second protein sequences, respectively, that are encoded by
10 DNA sequences representative of the same DNA population.

5. The method according to claim 2 in which the first
and second populations of fusion proteins comprise first and
second protein sequences, respectively, that are different.
15

6. The method according to claim 4 in which the first
and second populations of fusion proteins are each expressed
from chimeric genes comprising cDNA sequences of total
mammalian RNA or polyA' RNA of a cell.
20

7. The method according to claim 5 in which the first
and second populations of fusion proteins are each expressed
from chimeric genes comprising cDNA sequences of mammalian
RNA, and the first population of first fusion proteins is
25 expressed from chimeric genes comprising cDNA sequences of
diseased human tissue, and the second population of second
fusion proteins is expressed from chimeric genes comprising
cDNA sequences of non-diseased human tissue.

30   8. The method according to claim 6 in which the cDNA
sequences are of diseased human tissue.

9. The method according to claim 2 in which said first
or second population of fusion proteins has a complexity of
35 at least 10,000.

- 323 -

10.   The method according to claim 2 in which said first
or second population of fusion proteins has a complexity of
at least 50,000.

5   11.   The method according to claim 2 in which said first
and second populations of fusion proteins each has a
complexity of at least 10,000.

12.   The method according to claim 2 in which said first
10 and second populations of fusion proteins each has a
complexity of at least 50,000.

13.   The method according to claim 1 in which said first
population of first fusion proteins is expressed from a first
15 plasmid expression vector that expresses a first selectable
marker, and the second population of second fusion proteins
is expressed from a second plasmid expression vector that
expresses a second selectable marker, and in which the
population of host cells is incubated in an environment in
20 which substantial death of host cells occurs in the absence
of expression of the first and second selectable markers.

14.   The method according to claim 2 in which the
population of host cells is a population of mammalian host
25 cells.

15.   The method according to claim 2 in which the
population of host cells is a population of yeast host cells.

30   16.   The method according to claim 2 in which the
population of host cells is a population of bacterial host
cells.

17.   A method of detecting an inhibitor of a protein-
35 protein interaction comprising
        (a)   incubating a population of cells, said
    population comprising cells recombinantly expressing a

- 324 -

pair of interacting proteins, said pair consisting of a
first protein and a second protein, in the presence of
one or more candidate molecules among which it is
desired to identify an inhibitor of the interaction

5   between said first protein and said second protein, in
an environment in which substantial death of said cells
occurs (i) when said first protein and second protein
interact, or (ii) if said cells lack a recombinant
nucleic acid encoding said first protein or a

10   recombinant nucleic acid encoding said second protein;
and

        (b)   detecting those cells that survive said
incubating step, thereby detecting the presence of an
inhibitor of said interaction in said cells.

15

        18.   The method according to claim 17 in which the cells
are yeast cells.

        19.   The method according to claim 17 in which the first
20 protein and the second protein are first and second fusion
proteins, respectively, between which an interaction is
detected according to the method of claim 1.

        20.   The method according to claim 17 in which the first
25 protein and the second protein are first and second fusion
proteins, respectively, between which an interaction is
detected according to the method of claim 2.

        21.   A method of detecting one or more protein-protein
30 interactions comprising
        (a)   recombinantly expressing in a first population
of yeast cells of a first mating type, a first
population of first fusion proteins, each first fusion
protein comprising a first protein sequence and a DNA

35   binding domain, in which the DNA binding domain is the
same in each said first fusion protein; wherein said
first population of yeast cells contains a first

- 325 -

nucleotide sequence operably linked to a promoter driven
by one or more DNA binding sites recognized by said DNA
binding domain such that an interaction of a first
fusion protein with a second fusion protein, said second

5   fusion protein comprising a transcriptional activation
domain, results in increased transcription of said first
nucleotide sequence, and in which said first population
of first fusion proteins has a complexity of at least
1,000;

10   (b) negatively selecting to reduce the number of
those yeast cells expressing said first population of
first fusion proteins in which said increased
transcription of said first nucleotide sequence occurs
in the absence of said second fusion protein;

15   (c) recombinantly expressing in a second
population of yeast cells of a second mating type
different from said first mating type, a second
population of said second fusion proteins, each second
fusion protein comprising a second protein sequence and

20   an activation domain of a transcriptional activator, in
which the activation domain is the same in each said
second fusion protein, and in which said second
population of second fusion proteins has a complexity of
at least 1,000;

25   (d) mating said first population of yeast cells
with said second population of yeast cells to form a
population of diploid yeast cells, wherein said
population of diploid yeast cells contains a second
nucleotide sequence operably linked to a promoter driven

30   by a DNA binding site recognized by said DNA binding
domain such that an interaction of a first fusion
protein with a second fusion protein results in
increased transcription of said second nucleotide
sequence, in which the first and second nucleotide

35   sequences can be the same or different; and
          (e) detecting said increased transcription of said
first and/or second nucleotide sequence, thereby

- 326 -

detecting an interaction between a first fusion protein
and a second fusion protein.

22.   The method according to claim 21 in which said
5 negatively selecting is carried out by a method comprising
incubating said first population of yeast cells expressing
said first population of first fusion proteins in an
environment in which substantial death of said first
population of host cells occurs if said increased
10 transcription occurs.

23.   The method according to claim 22 in which said
first nucleotide sequence comprises a functional *URA3* coding
sequence, and said environment contains 5-fluoroorotic acid.
15

24.   The method according to claim 22 in which said
first nucleotide sequence comprises a functional *LYS2* coding
sequence, and said environment comprises α-amino-adipate.

20   25.   The method according to claim 21 in which said
second nucleotide sequence is different from said first
nucleotide sequence.

26.   The method according to claim 23 in which said
25 second nucleotide sequence comprises a functional *lacZ* coding
sequence.

27.   The method according to claim 21 in which said
first and second nucleotide sequences are selected from the
30 group consisting of the functional coding sequences of *URA3*,
*HIS3, lacZ, GFP, LEU2, LYS2, ADE2, TRP1, CAN1, CYH2, GUS,
CUP1*, and *CAT*.

28.   The method according to claim 21 in which said DNA
35 binding domain is selected from the group consisting of the
DNA binding domains of GAL4, GCN4, ARD1, LEX A, and Ace1N.

29. The method according to claim 21 in which said transcription activation domain is selected from the group consisting of the activation domains of GAL4, GCN4, ARD1, herpes simplex virus VP16, and Ace1C.

30. The method according to claim 21 in which said first population of yeast cells and said second population of yeast cells do not contain functional counterparts of said first and second nucleotide sequences that are not operably linked to a promoter driven by said one or more DNA binding sites.

31. The method according to claim 21 in which said DNA binding domain is a GAL4 or LEX A DNA binding domain, and said transcription activation domain is a GAL4 or herpes simplex virus VP16 activation domain.

32. The method according to claim 21 in which the first and second populations of fusion proteins are each expressed from chimeric genes comprising cDNA sequences from an uncharacterized sample of a population of cDNA from mammalian RNA.

33. The method according to claim 21 in which the first and second populations of fusion proteins comprise first and second protein sequences, respectively, that are encoded by DNA sequences representative of the same DNA population.

34. The method according to claim 21 in which the first and second populations of fusion proteins comprise first and second protein sequences, respectively, that are different.

35. The method according to claim 33 in which the first and second populations of fusion proteins are each expressed from chimeric genes comprising cDNA sequences of mammalian RNA.

36. The method according to claim 34 in which the first and second populations of fusion proteins are each expressed from chimeric genes comprising cDNA sequences of mammalian RNA, and the first population of first fusion proteins is

5 expressed from chimeric genes comprising cDNA sequences of diseased human tissue, and the second population of second fusion proteins is expressed from chimeric genes comprising cDNA sequences of non-diseased human tissue.

10   37. The method according to claim 35 in which the cDNA sequences are of diseased human tissue.

38. The method according to claim 21 in which said first or second population of fusion proteins has a

15 complexity of at least 10,000.

39. The method according to claim 21 in which said first or second population of fusion proteins has a complexity of at least 50,000.

20

40. The method according to claim 21 in which said first and second populations of fusion proteins each has a complexity of at least 10,000.

25   41. The method according to claim 21 in which said first and second populations of fusion proteins each has a complexity of at least 50,000.

42. The method according to claim 21, 22 or 23 in which

30 in said mating step at least 5.8×10^8 matings are done.

43. The method according to claim 21, 22 or 23 in which in said mating step at least 8.5×10^9 matings are done.

35   44. The method according to claim 21, 22 or 23 in which in said mating step at least 8.5×10^{10} matings are done.

45. The method according to claim 21, 22 or 23 in which said mating is performed on solid medium.

46. The method according to claim 21 in which the first
5 population of first fusion proteins is expressed from a first plasmid expression vector that expresses a first selectable marker, and the second population of second fusion proteins is expressed from a second plasmid expression vector that expresses a second selectable marker, and in which the first
10 population of yeast cells is incubated in a first environment in which substantial death of yeast cells occurs in the absence of expression of the first selectable marker, and the second population of yeast cells is incubated in a second environment in which substantial death of yeast cells occurs
15 in the absence of expression of the second selectable marker.

47. The method according to claim 21 in which the yeast cells are *Saccharomyces cerevisiae*.

20 48. The method according to claim 17 in which the cells are yeast cells, and in which the first protein and the second protein are first and second fusion proteins, respectively, between which an interaction is detected according to the method of claim 21.
25

49. The method according to claim 48 in which the yeast cells contain functional *URA3* coding sequences under the control of a promoter driven by a DNA binding site recognized by said DNA binding domain of said first fusion protein, and
30 said environment contains 5-fluoroorotic acid.

50. The method according to claim 48 or 49 in which said one or more candidate molecules are provided to said at least one diploid yeast cell by introducing into said diploid
35 yeast cell one or more recombinant nucleic acids encoding said one or more candidate molecules, such that said one or

more candidate molecules are expressed within said diploid
yeast cell.

    51.   The method according to claim 48 or 49 in which
5 said one or more candidate molecules are provided to said at
least one diploid yeast cell by incubating said diploid yeast
cell in an environment comprising said one or more candidate
molecules.

10   52.   A method of detecting one or more protein-protein
interactions present within a first protein population and
absent within a second protein population comprising
      (a)   carrying out the method of claim 1 wherein
said first protein sequences of said first fusion
15   proteins and said second protein sequences of said
second fusion proteins are encoded by DNA sequences
representative of the same first DNA population, thereby
detecting one or more protein-protein interactions;
      (b)   carrying out the method of claim 1 wherein
20   said first protein sequences of said first fusion
proteins and said second protein sequences of said
second fusion proteins are encoded by DNA sequences
representative of the same second DNA population, said
second DNA population differing from said first DNA
25   population, thereby detecting one or more protein-
protein interactions; and
      (c)   comparing the one or more protein-protein
interactions detected in step (a) with the one or more
protein-protein interactions detected in step (b).
30

    53.   A method of detecting one or more protein-protein
interactions present within a first protein population and
absent within a second protein population comprising
      (a)   carrying out the method of claim 21 wherein
35   said first protein sequences of said first fusion
proteins and said second protein sequences of said
second fusion proteins are encoded by DNA sequences

representative of the same first DNA population, thereby detecting one or more protein-protein interactions;

(b)   carrying out the method of claim 21 wherein said first protein and said second protein sequences of
5   said second fusion proteins are encoded by DNA sequences representative of the same second DNA population, said second DNA population differing from said first DNA population, thereby detecting one or more protein-protein interactions; and
10   (c)   comparing the one or more protein-protein interactions detected in step (a) with the one or more protein-protein interactions detected in step (b).

54.   A method of detecting one or more protein-protein
15 interactions comprising

(a)   introducing into a first population of cells of *Saccharomyces cerevisiae* a first population of first plasmids, each said first plasmid encoding and capable of expressing in the first population of cells (i) *TRP1*,
20   and (ii) a first population of first fusion proteins, each said first fusion protein comprising a GAL4 DNA binding domain and a first protein sequence, in which said first population of first fusion proteins has a complexity of at least 1,000, and in which said first
25   population of cells (i) is of a first mating type selected from the group consisting of a and α, (ii) is mutant in endogenous *URA3* and *HIS3*, (iii) contains functional *URA3* coding sequences under the control of a promoter containing GAL4 binding sites, and (iv)
30   contains functional *lacZ* coding sequences under the control of a promoter containing GAL4 binding sites;

(b)   introducing into a second population of cells of *Saccharomyces cerevisiae* a second population of second plasmids, each said second plasmids encoding and
35   capable of expressing in the second population of cells (i) *LEU2*, and (ii) a second population of second fusion proteins, each said second fusion protein comprising a

GAL4 transcriptional activation domain and a second
protein sequence, in which said second population of
second fusion proteins has a complexity of at least
1,000, and in which said second population of cells (i)
is of a second mating type different from said first
mating type and selected from the group consisting of a
and α, (ii) is mutant in endogenous *URA3* and *HIS3*, (iii)
contains functional *HIS3* coding sequences under the
control of a promoter containing GAL4 binding sites, and
(iv) contains functional *lacZ* coding sequences under the
control of a promoter containing GAL4 binding sites;

(c)   after step (a), incubating said first
population of cells in an environment lacking tryptophan
and containing 5-fluoroorotic acid;

(d)   pooling surviving cells from said first
population after step (c);

(e)   after step (b), incubating said second
population of cells in an environment lacking leucine;

(f)   pooling surviving cells from said second
population after step (e);

(g)   mating the pooled cells from said first
population and the pooled cells from said second
population by mixing the cells together, applying the
cells to a solid medium and incubating the cells, to
form diploid cells; and

(h)   incubating the diploid cells in an environment
lacking uracil, histidine, tryptophan and leucine, to
select diploid cells containing a said first plasmid and
a said second plasmid and in which transcription of the
URA3 and *HIS3* coding sequences has been activated,
thereby indicating that a first fusion protein has
interacted with a second fusion protein within the
diploid cell, thereby detecting one or more protein-
protein interactions.

55.   The method according to claim 54 in which the
pooled cells from said first population and the pooled cells

- 333 -

from said second population in said mating step are each at
least 8.5 x 10^9 in number.

56.   The method according to claim 54 in which the first
5 and second populations of fusion proteins are each expressed
from chimeric genes comprising cDNA sequences of mammalian
RNA.

57.   The method according to claim 1 which further
10 comprises obtaining a purified DNA encoding said first fusion
protein or encoding a portion thereof comprising said first
protein sequence, from a host cell in which said regulation
of transcription is detected.

15   58.   The method according to claim 1 which further
comprises obtaining a purified DNA encoding said second
fusion protein or encoding a portion thereof comprising said
second protein sequence, from a host cell in which said
regulation of transcription is detected.
20

59.   The method according to claim 21 which further
comprises obtaining a purified DNA encoding said first fusion
protein or encoding a portion thereof comprising said first
protein sequence, from a yeast cell in which said increased
25 transcription is detected in step (e).

60.   The method according to claim 21 which further
comprises obtaining a purified DNA encoding said second
fusion protein or encoding a portion thereof comprising said
30 second protein sequence, from a yeast cell in which said
increased transcription is detected in step (e).

61.   The method according to claim 46 in which said
first and second plasmid expression vectors are replicable
35 both in yeast cells and in *E. coli*.

62.   The method according to claim 54 in which said
first plasmids and said second plasmids are replicable both
in yeast cells and in *E. coli*.

5   63.   The method according to claim 54 which further
comprises obtaining a first purified DNA encoding said first
fusion protein or encoding a portion thereof comprising said
first protein sequence, from a yeast cell selected in step
(h).

10

64.   The method according to claim 54 which further
comprises obtaining a purified DNA encoding said second
fusion protein or encoding a portion thereof comprising said
second protein sequence, from a yeast cell selected in step
15 (h).

65.   The method according to claim 57, 59, or 63 which
further comprises sequencing at least a portion of said
purified DNA to determine the sequence of said first protein
20 sequence.

66.   The method according to claim 58, 60, or 64 which
further comprises sequencing at least a portion of said
purified DNA to determine the sequence of said second protein
25 sequence.

67.   The method according to claim 1 which further
comprises amplifying of DNA fragments encoding said first
fusion protein or encoding a portion thereof comprising said
30 first protein sequence from a plurality of host cells in
which said regulation of transcription is detected, and
subjecting a sample comprising said resulting amplified DNA
fragments to a method for identifying, classifying, or
quantifying one or more nucleic acids in the sample, said
35 method comprising:

        (a) probing said sample with one or more
        recognition means, each recognition means causing

recognition of a target nucleotide subsequence or a set
of target nucleotide subsequences;

       (b) generating one or more signals from said sample
probed by said recognition means, each generated signal
5   arising from a nucleic acid in said sample and
comprising a representation of (i) the identities of
effective subsequences, each said effective subsequence
being a subsequence comprising a target subsequence, or
the identities of sets of effective subsequences, each
10   said set having member effective subsequences each of
which comprises a different target subsequence from one
of said sets of target sequences, and (ii) the length
between occurrences of effective subsequences in said
nucleic acid or between one occurrence of one effective
15   subsequence and the end of said nucleic acid; and

       (c) searching a nucleotide sequence database to
determine sequences that match or the absence of any
sequences that match said one or more generated signals,
said database comprising a plurality of known nucleotide
20   sequences of nucleic acids that may be present in the
sample, a sequence from said database matching a
generated signal when the sequence from said database
has both (i) the same length between occurrences of
effective subsequences or the same length between one
25   occurrence of one effective target subsequence and the
end of the sequence as is represented by the generated
signal, and (ii) the same effective subsequences as are
represented by the generated signal, or effective
subsequences that are members of the same sets of
30   effective subsequences as are represented by the
generated signal,
whereby said one or more nucleic acids in said sample are
identified, classified, or quantified.

35   68. The method according to claim 21 which further
comprises amplifying of DNA fragments encoding said first
fusion protein or encoding a portion thereof comprising said

first protein sequence from a plurality of yeast cells in
which said increased transcription is detected in step (e),
and subjecting a sample comprising said resulting amplified
DNA fragments to a method for identifying, classifying, or
5 quantifying one or more nucleic acids in the sample, said
method comprising:

(a) probing said sample with one or more
recognition means, each recognition means causing
recognition of a target nucleotide subsequence or a set
10   of target nucleotide subsequences;

(b) generating one or more signals from said sample
probed by said recognition means, each generated signal
arising from a nucleic acid in said sample and
comprising a representation of (i) the identities of
15   effective subsequences, each said effective subsequence
being a subsequence comprising a target subsequence, or
the identities of sets of effective subsequences, each
said set having member effective subsequences each of
which comprises a different target subsequence from one
20   of said sets of target sequences, and (ii) the length
between occurrences of effective subsequences in said
nucleic acid or between one occurrence of one effective
subsequence and the end of said nucleic acid; and

(c) searching a nucleotide sequence database to
25   determine sequences that match or the absence of any
sequences that match said one or more generated signals,
said database comprising a plurality of known nucleotide
sequences of nucleic acids that may be present in the
sample, a sequence from said database matching a
30   generated signal when the sequence from said database
has both (i) the same length between occurrences of
effective subsequences or the same length between one
occurrence of one effective target subsequence and the
end of the sequence as is represented by the generated
35   signal, and (ii) the same effective subsequences as are
represented by the generated signal, or effective
subsequences that are members of the same sets of

effective subsequences as are represented by the
generated signal,
whereby said one or more nucleic acids in said sample are
identified, classified, or quantified.

5

69. The method according to claim 54 which further
comprises amplifying of DNA fragments encoding said first
fusion protein or encoding a portion thereof comprising said
first protein sequence from a plurality of cells which
10 survive said incubating step (h), and subjecting a sample
comprising said resulting amplified fragments to a method for
identifying, classifying, or quantifying one or more nucleic
acids in the sample, said method comprising:

      (a) probing said sample with one or more
15 recognition means, each recognition means causing
recognition of a target nucleotide subsequence or a set
of target nucleotide subsequences;

      (b) generating one or more signals from said sample
probed by said recognition means, each generated signal
20 arising from a nucleic acid in said sample and
comprising a representation of (i) the identities of
effective subsequences, each said effective subsequence
being a subsequence comprising a target subsequence, or
the identities of sets of effective subsequences, each
25 said set having member effective subsequences each of
which comprises a different target subsequence from one
of said sets of target sequences, and (ii) the length
between occurrences of effective subsequences in said
nucleic acid or between one occurrence of one effective
30 subsequence and the end of said nucleic acid; and

      (c) searching a nucleotide sequence database to
determine sequences that match or the absence of any
sequences that match said one or more generated signals,
said database comprising a plurality of known nucleotide
35 sequences of nucleic acids that may be present in the
sample, a sequence from said database matching a
generated signal when the sequence from said database

- 338 -

has both (i) the same length between occurrences of
effective subsequences or the same length between one
occurrence of one effective target subsequence and the
end of the sequence as is represented by the generated
5   signal, and (ii) the same effective subsequences as are
represented by the generated signal, or effective
subsequences that are members of the same sets of
effective subsequences as are represented by the
generated signal,
10 whereby said one or more nucleic acids in said sample are
identified, classified, or quantified.

70.   The method according to claim 1 which further
comprises amplifying of DNA fragments encoding said second
15 fusion protein or encoding a portion thereof comprising said
second protein sequence from a plurality of host cells in
which said regulation of transcription is detected, and
subjecting a sample comprising said resulting amplified DNA
fragments to a method for identifying, classifying, or
20 quantifying one or more nucleic acids in the sample, said
method comprising:
        (a) probing said sample with one or more
recognition means, each recognition means causing
recognition of a target nucleotide subsequence or a set
25   of target nucleotide subsequences;
        (b) generating one or more signals from said sample
probed by said recognition means, each generated signal
arising from a nucleic acid in said sample and
comprising a representation of (i) the identities of
30   effective subsequences, each said effective subsequence
being a subsequence comprising a target subsequence, or
the identities of sets of effective subsequences, each
said set having member effective subsequences each of
which comprises a different target subsequence from one
35   of said sets of target sequences, and (ii) the length
between occurrences of effective subsequences in said

- 339 -

nucleic acid or between one occurrence of one effective
subsequence and the end of said nucleic acid; and

(c) searching a nucleotide sequence database to
determine sequences that match or the absence of any
5   sequences that match said one or more generated signals,
said database comprising a plurality of known nucleotide
sequences of nucleic acids that may be present in the
sample, a sequence from said database matching a
generated signal when the sequence from said database
10   has both (i) the same length between occurrences of
effective subsequences or the same length between one
occurrence of one effective target subsequence and the
end of the sequence as is represented by the generated
signal, and (ii) the same effective subsequences as are
15   represented by the generated signal, or effective
subsequences that are members of the same sets of
effective subsequences as are represented by the
generated signal,
whereby said one or more nucleic acids in said sample are
20 identified, classified, or quantified.

71.   The method according to claim 21 which further
comprises amplifying of DNA fragments encoding said second
fusion protein or encoding a portion thereof comprising said
25 second protein sequence from a plurality of yeast cells in
which said increased transcription is detected in step (e),
and subjecting a sample comprising said resulting amplified
DNA fragments to a method for identifying, classifying, or
quantifying one or more nucleic acids in the sample, said
30 method comprising:

(a) probing said sample with one or more
recognition means, each recognition means causing
recognition of a target nucleotide subsequence or a set
of target nucleotide subsequences;
35   (b) generating one or more signals from said sample
probed by said recognition means, each generated signal
arising from a nucleic acid in said sample and

- 340 -

comprising a representation of (i) the identities of
effective subsequences, each said effective subsequence
being a subsequence comprising a target subsequence, or
the identities of sets of effective subsequences, each

5   said set having member effective subsequences each of
which comprises a different target subsequence from one
of said sets of target sequences, and (ii) the length
between occurrences of effective subsequences in said
nucleic acid or between one occurrence of one effective

10   subsequence and the end of said nucleic acid; and

        (c) searching a nucleotide sequence database to
determine sequences that match or the absence of any
sequences that match said one or more generated signals,
said database comprising a plurality of known nucleotide

15   sequences of nucleic acids that may be present in the
sample, a sequence from said database matching a
generated signal when the sequence from said database
has both (i) the same length between occurrences of
effective subsequences or the same length between one

20   occurrence of one effective target subsequence and the
end of the sequence as is represented by the generated
signal, and (ii) the same effective subsequences as are
represented by the generated signal, or effective
subsequences that are members of the same sets of

25   effective subsequences as are represented by the
generated signal,
whereby said one or more nucleic acids in said sample are
identified, classified, or quantified.

30   72. The method according to claim 54 which further
comprises amplifying of DNA fragments encoding said second
fusion protein or encoding a portion thereof comprising said
second protein sequence from a plurality of cells which
survive said incubating step (h), and subjecting a sample
35 comprising said resulting amplified fragments to a method for
identifying, classifying, or quantifying one or more nucleic
acids in the sample, said method comprising:

- 341 -

(a) probing said sample with one or more recognition means, each recognition means causing recognition of a target nucleotide subsequence or a set of target nucleotide subsequences;

5   (b) generating one or more signals from said sample probed by said recognition means, each generated signal arising from a nucleic acid in said sample and comprising a representation of (i) the identities of effective subsequences, each said effective subsequence

10   being a subsequence comprising a target subsequence, or the identities of sets of effective subsequences, each said set having member effective subsequences each of which comprises a different target subsequence from one of said sets of target sequences, and (ii) the length

15   between occurrences of effective subsequences in said nucleic acid or between one occurrence of one effective subsequence and the end of said nucleic acid; and

      (c) searching a nucleotide sequence database to determine sequences that match or the absence of any

20   sequences that match said one or more generated signals, said database comprising a plurality of known nucleotide sequences of nucleic acids that may be present in the sample, a sequence from said database matching a generated signal when the sequence from said database

25   has both (i) the same length between occurrences of effective subsequences or the same length between one occurrence of one effective target subsequence and the end of the sequence as is represented by the generated signal, and (ii) the same effective subsequences as are

30   represented by the generated signal, or effective subsequences that are members of the same sets of effective subsequences as are represented by the generated signal,

   whereby said one or more nucleic acids in said sample are

35 identified, classified, or quantified.

73.   The method according to claim 1 which further comprises amplifying of DNA fragments encoding said first fusion protein or encoding a portion thereof comprising said first protein sequence from a plurality of host cells in

5 which said regulation of transcription is detected, and subjecting a sample comprising said resulting amplified DNA fragments to a method for identifying, classifying, or quantifying DNA molecules in the sample, said method comprising:

10   (a) digesting said sample with one or more restriction endonucleases, each said restriction endonuclease recognizing a subsequence recognition site and digesting DNA at said recognition site to produce fragments with 5' overhangs;

15   (b) contacting said produced fragments with shorter and longer oligodeoxynucleotides, each said shorter oligodeoxynucleotide hybridizable with a said 5' overhang and having no terminal phosphates, each said longer oligodeoxynucleotide hybridizable with a said

20   shorter oligodeoxynucleotide;

          (c) ligating said longer oligodeoxynucleotides to said 5' overhangs on said fragments to produce ligated DNA fragments;

          (d) extending said ligated DNA fragments by

25   synthesis with a DNA polymerase to produce blunt-ended double stranded DNA fragments;

          (e) amplifying said blunt-ended double stranded DNA fragments by a method comprising contacting said blunt-ended double stranded DNA fragments with a DNA

30   polymerase and primer oligodeoxynucleotides, each said primer oligodeoxynucleotide having a sequence comprising that of one of the longer oligodeoxynucleotides;

          (f) determining the length of the amplified DNA fragments produced in step (e); and

35   (g) searching a DNA sequence database, said database comprising a plurality of known DNA sequences that may be present in the sample, for sequences

- 343 -

matching one or more of said fragments of determined
length, a sequence from said database matching a
fragment of determined length when the sequence from
said database comprises recognition sites of said one or
5   more restriction endonucleases spaced apart by the
determined length,
whereby DNA molecules in said sample are identified,
classified, or quantified.

10   74. The method according to claim 1 which further
comprises amplifying of DNA fragments encoding said second
fusion protein or encoding a portion thereof comprising said
second protein sequence from a plurality of host cells in
which said regulation of transcription is detected, and
15 subjecting a sample comprising said resulting amplified DNA
fragments to a method for identifying, classifying, or
quantifying DNA molecules in the sample, said method
comprising:
        (a) digesting said sample with one or more
20   restriction endonucleases, each said restriction
endonuclease recognizing a subsequence recognition site
and digesting DNA at said recognition site to produce
fragments with 5' overhangs;
        (b)   contacting said produced fragments with
25   shorter and longer oligodeoxynucleotides, each said
shorter oligodeoxynucleotide hybridizable with a said 5'
overhang and having no terminal phosphates, each said
longer oligodeoxynucleotide hybridizable with a said
shorter oligodeoxynucleotide;
30   (c)   ligating said longer oligodeoxynucleotides to
said 5' overhangs on said fragments to produce ligated
DNA fragments;
        (d)   extending said ligated DNA fragments by
synthesis with a DNA polymerase to produce blunt-ended
35   double stranded DNA fragments;
        (e)   amplifying said blunt-ended double stranded
DNA fragments by a method comprising contacting said

- 344 -

blunt-ended double stranded DNA fragments with a DNA
polymerase and primer oligodeoxynucleotides, each said
primer oligodeoxynucleotide having a sequence comprising
that of one of the longer oligodeoxynucleotides;

5   (f)   determining the length of the amplified DNA
fragments produced in step (e); and

           (g)   searching a DNA sequence database, said
database comprising a plurality of known DNA sequences
that may be present in the sample, for sequences

10   matching one or more of said fragments of determined
length, a sequence from said database matching a
fragment of determined length when the sequence from
said database comprises recognition sites of said one or
more restriction endonucleases spaced apart by the

15   determined length,
whereby DNA molecules in said sample are identified,
classified, or quantified.

       75.   A method of determining one or more characteristics
20 of or the identities of nucleic acids encoding an interacting
pair of proteins from among a population of cells containing
a multiplicity of different nucleic acids encoding different
pairs of interacting proteins, said method comprising:

           (a) designating each group of cells containing
25   nucleic acids encoding an identical pair of interacting
proteins as one point of a multidimensional array in
which the intersection of axes in each dimension
uniquely identifies a single said group;

           (b)   pooling all groups along a simple axis to form
30   a plurality of pooled groups;

           (c)   amplifying from a first aliquot of each pooled
group a plurality of first nucleic acids, each first
nucleic acid comprising a sequence encoding a first
protein that is one-half of a pair of interacting

35   proteins;

           (d)   amplifying from a second aliquot of each
pooled group a plurality of second nucleic acids, each

second nucleic acid comprising a sequence encoding a
second protein that is the other half of the pair of
interacting proteins;

     (e)   subjecting said first nucleic acids from each
5   pooled group to size separation;

     (f)   subjecting said second nucleic acids from each
pooled group to size separation;

     (g)   identifying which at least one of said first
nucleic acids are present in samples of first nucleic
10   acids from a pooled group from each axes in each
dimension, thereby indicating that said at least one
first nucleic acid is present in said array in the group
designated at the intersection of said axes in each
dimension; and

15   (h)   identifying which at least one of said second
nucleic acids are present in samples of a second nucleic
acid from a pooled group from axes in each dimension,
thereby indicating that the said at least one second
nucleic acid is present in said array in the group
20   designated at the intersection of said axes in each
dimension;

in which the first and second nucleic acids that are
indicated to be present in said array in a group designated
at the same intersection are indicated to encode interacting
25 proteins.

    76. A method of determining one or more characteristics
of or the identities of nucleic acids encoding an interacting
pair of proteins from among a plurality of yeast cell
30 colonies, each colony containing nucleic acids encoding a
different pair of interacting proteins, said method
comprising carrying out the method of claim 21 in which an
interaction between a first fusion protein and a second
fusion protein is detected in a plurality of colonies of
35 diploid yeast cells, and which method further comprises:

     (f)   designating each colony in which an
interaction between a first fusion protein and a second

- 346 -

fusion protein is detected as one point of a
multidimensional array in which the intersection of axes
in each dimension uniquely identifies a single said
colony;

5   (g)   pooling all colonies along a simple axis to
form a plurality of pooled colonies;

          (h)   amplifying from a first aliquot of each pooled
colony a plurality of first nucleic acids, each first
nucleic acid comprising a sequence encoding said first
10   fusion protein or a portion thereof comprising said
first protein sequence;

          (i)   amplifying from a second aliquot of each
pooled colony a plurality of second nucleic acids, each
second nucleic acid comprising a sequence encoding said
15   second fusion protein or a portion thereof comprising
said second protein sequence;

          (j)   subjecting said first nucleic acids from each
pooled colony to size separation;

          (k)   subjecting said second nucleic acids from each
20   pooled colony to size separation;

          (l)   identifying which at least one of said first
nucleic acids are present in samples of first nucleic
acids from a pooled colony from axes in each dimension,
thereby indicating that said at least one first nucleic
25   acid is present in said array in the colony designated
at the intersection of said axes in each dimension;

          (m)   identifying which at least one of said second
nucleic acids are present in samples of a second nucleic
acid from a pooled colony from axes in each dimension,
30   thereby indicating that the said at least one second
nucleic acid is present in said array in the colony
designated at the intersection of said axes in each
dimension;

in which the first and second nucleic acids that are
35 indicated to be present in said array in a colony designated
at the same intersection are indicated to encode interacting
protein sequences.

77.   A method of determining one or more characteristics
of or the identities of DNA molecules encoding an interacting
pair of proteins from among a plurality of yeast cell
colonies, each colony containing DNA molecules encoding a
5 different pair of interacting proteins, comprising carrying
out the method of claim 54 in which an interaction between a
first fusion protein and a second fusion protein is detected
in a plurality of colonies of diploid yeast cells, and which
method further comprises:

10   (f)   designating each colony in which an
interaction between a first fusion protein and a second
fusion protein is detected as one point of a
multidimensional array in which the intersection of axes
in each dimension uniquely identifies a single said
15   colony;

           (g)   pooling all colonies along a simple axis to
form a plurality of pooled colonies;

           (h)   amplifying from a first aliquot of each pooled
colony a plurality of first DNA molecules, each first
20   DNA molecule comprising a sequence encoding said first
fusion protein or a portion thereof comprising said
first protein sequence;

           (i)   amplifying from a second aliquot of each
pooled colony a plurality of second DNA molecules, each
25   second DNA molecule comprising a sequence encoding said
second fusion protein or a portion thereof comprising
said second protein sequence;

           (j)   subjecting said first DNA molecules from each
pooled colony to size separation;

30   (k)   subjecting said second DNA molecules from each
pooled colony to size separation;

           (l)   identifying which at least one of said first
DNA molecules are present in samples of first DNA
molecules from a pooled colony from axes in each
35   dimension, thereby indicating that said at least one
first DNA molecule is present in said array in the

- 348 -

colony designated at the intersection of said axes in
each dimension;

     (m) identifying which at least one of said second
DNA molecules are present in samples of a second DNA
5   molecule from a pooled colony from axes in each
dimension, thereby indicating that the said at least one
second DNA molecule is present in said array in the
colony designated at the intersection of said axes in
each dimension;

10 in which the first and second DNA molecules that are
indicated to be present in said array in a colony designated
at the same intersection are indicated to encode interacting
protein sequences.

15   78. The method according to claim 76 or 77 in which
said amplifying is by use of polymerase chain reaction.

    79. The method according to claim 76 which further
comprises determining the nucleotide sequence of at least one
20 first nucleic acid or second nucleic acid.

    80. The method according to claim 77 which further
comprises determining the nucleotide sequence of at least one
first DNA molecule or second DNA molecule.

25

    81. The method according to claim 76 which further
comprises subjecting said pooled colonies of first nucleic
acids to a method for identifying, classifying, or
quantifying one or more nucleic acids in a sample, said
30 method comprising:

    (a) probing said sample with one or more
recognition means, each recognition means causing
recognition of a target nucleotide subsequence or a set
of target nucleotide subsequences;

35   (b) generating one or more signals from said sample
probed by said recognition means, each generated signal
arising from a nucleic acid in said sample and

- 349 -

comprising a representation of (i) the identities of
effective subsequences, each said effective subsequence
being a subsequence comprising a target subsequence, or
the identities of sets of effective subsequences, each
5   said set having member effective subsequences each of
which comprises a different target subsequence from one
of said sets of target sequences, and (ii) the length
between occurrences of effective subsequences in said
nucleic acid or between one occurrence of one effective
10 subsequence and the end of said nucleic acid; and
        (c) searching a nucleotide sequence database to
determine sequences that match or the absence of any
sequences that match said one or more generated signals,
said database comprising a plurality of known nucleotide
15 sequences of nucleic acids that may be present in the
sample, a sequence from said database matching a
generated signal when the sequence from said database
has both (i) the same length between occurrences of
effective subsequences or the same length between one
20 occurrence of one effective target subsequence and the
end of the sequence as is represented by the generated
signal, and (ii) the same effective subsequences as are
represented by the generated signal, or effective
subsequences that are members of the same sets of
25 effective subsequences as are represented by the
generated signal,
whereby said one or more nucleic acids in said sample are
identified, classified, or quantified.

30 82. The method according to claim 77 which further
comprises subjecting said pooled colonies of first DNA
molecules to a method for identifying, classifying, or
quantifying one or more DNA molecules in a sample, said
method comprising:
35   (a) probing said sample with one or more
recognition means, each recognition means causing

- 350 -

recognition of a target nucleotide subsequence or a set
of target nucleotide subsequences;

(b) generating one or more signals from said sample
probed by said recognition means, each generated signal
arising from a nucleic acid in said sample and
comprising a representation of (i) the identities of
effective subsequences, each said effective subsequence
being a subsequence comprising a target subsequence, or
the identities of sets of effective subsequences, each
said set having member effective subsequences each of
which comprises a different target subsequence from one
of said sets of target sequences, and (ii) the length
between occurrences of effective subsequences in said
nucleic acid or between one occurrence of one effective
subsequence and the end of said nucleic acid; and

(c) searching a nucleotide sequence database to
determine sequences that match or the absence of any
sequences that match said one or more generated signals,
said database comprising a plurality of known nucleotide
sequences of nucleic acids that may be present in the
sample, a sequence from said database matching a
generated signal when the sequence from said database
has both (i) the same length between occurrences of
effective subsequences or the same length between one
occurrence of one effective target subsequence and the
end of the sequence as is represented by the generated
signal, and (ii) the same effective subsequences as are
represented by the generated signal, or effective
subsequences that are members of the same sets of
effective subsequences as are represented by the
generated signal,
whereby said one or more nucleic acids in said sample are
identified, classified, or quantified.

83.   The method according to claim 76 which further
comprises subjecting said pooled colonies of second nucleic
acids to a method comprising a method for identifying,

classifying, or quantifying one or more nucleic acids in a
sample, said method comprising:

       (a) probing said sample with one or more
recognition means, each recognition means causing

5   recognition of a target nucleotide subsequence or a set
of target nucleotide subsequences;

       (b) generating one or more signals from said sample
probed by said recognition means, each generated signal
arising from a nucleic acid in said sample and

10   comprising a representation of (i) the identities of
effective subsequences, each said effective subsequence
being a subsequence comprising a target subsequence, or
the identities of sets of effective subsequences, each
said set having member effective subsequences each of

15   which comprises a different target subsequence from one
of said sets of target sequences, and (ii) the length
between occurrences of effective subsequences in said
nucleic acid or between one occurrence of one effective
subsequence and the end of said nucleic acid; and

20   (c) searching a nucleotide sequence database to
determine sequences that match or the absence of any
sequences that match said one or more generated signals,
said database comprising a plurality of known nucleotide
sequences of nucleic acids that may be present in the

25   sample, a sequence from said database matching a
generated signal when the sequence from said database
has both (i) the same length between occurrences of
effective subsequences or the same length between one
occurrence of one effective target subsequence and the

30   end of the sequence as is represented by the generated
signal, and (ii) the same effective subsequences as are
represented by the generated signal, or effective
subsequences that are members of the same sets of
effective subsequences as are represented by the

35   generated signal,
whereby said one or more nucleic acids in said sample are
identified, classified, or quantified.

84. The method according to claim 77 which further
comprises subjecting said pooled colonies of second DNA
molecules to a method comprising a method for identifying,
classifying, or quantifying one or more DNA molecules in a
5 sample, said method comprising:

(a) probing said sample with one or more
recognition means, each recognition means causing
recognition of a target nucleotide subsequence or a set
of target nucleotide subsequences;

10 (b) generating one or more signals from said sample
probed by said recognition means, each generated signal
arising from a nucleic acid in said sample and
comprising a representation of (i) the identities of
effective subsequences, each said effective subsequence
15 being a subsequence comprising a target subsequence, or
the identities of sets of effective subsequences, each
said set having member effective subsequences each of
which comprises a different target subsequence from one
of said sets of target sequences, and (ii) the length
20 between occurrences of effective subsequences in said
nucleic acid or between one occurrence of one effective
subsequence and the end of said nucleic acid; and

(c) searching a nucleotide sequence database to
determine sequences that match or the absence of any
25 sequences that match said one or more generated signals,
said database comprising a plurality of known nucleotide
sequences of nucleic acids that may be present in the
sample, a sequence from said database matching a
generated signal when the sequence from said database
30 has both (i) the same length between occurrences of
effective subsequences or the same length between one
occurrence of one effective target subsequence and the
end of the sequence as is represented by the generated
signal, and (ii) the same effective subsequences as are
35 represented by the generated signal, or effective
subsequences that are members of the same sets of

effective subsequences as are represented by the
generated signal,
whereby said one or more nucleic acids in said sample are
identified, classified, or quantified.

5

85.   The method according to claim 46 in which the
population of diploid yeast cells is incubated in a third
environment in which substantial death of yeast cells occurs
in the absence of expression of the first and second
10 selectable markers.

86.   Purified cells of a single yeast strain of mating
type a, that is mutant in endogenous *URA3* and *HIS3*, and
contains functional *URA3* coding sequences under the control
15 of a promoter containing GAL4 binding sites, and contains
functional *lacZ* coding sequences under the control of a
promoter containing GAL4 binding sites.

87.   Purified cells of a single yeast strain of mating
20 type α, that is mutant in endogenous *URA3* and *HIS3*, and
contains functional *URA3* coding sequences under the control
of a promoter containing GAL4 binding sites, and contains
functional *lacZ* coding sequences under the control of a
promoter containing GAL4 binding sites.

25

88. A kit comprising in one or more containers:
        (a) purified cells of a single yeast strain of
        mating type a, that is mutant in endogenous *URA3* and
        HIS3, and contains functional *URA3* coding sequences
30   under the control of a promoter containing GAL4 binding
        sites, and contains functional *lacZ* coding sequences
        under the control of a promoter containing GAL4 binding
        sites; and
        (b) purified cells of a single yeast strain of
35   mating type α, that is mutant in endogenous *URA3* and
        HIS3, and contains functional *URA3* coding sequences
        under the control of a promoter containing GAL4 binding

- 354 -

sites, and contains functional *lacZ* coding sequences
under the control of a promoter containing GAL4 binding
sites.

5   89.   The kit of claim 88 which further comprises in one
or more containers:
           (c) a first vector comprising:
                 (i)   a first promoter;
                 (ii) a first nucleotide sequence encoding a
10   DNA binding domain, operably linked to the
                       first promoter;
                 (iii) means for inserting a DNA sequence
                       encoding a protein into the vector in such a
                       manner that the protein is capable of being
15   expressed as part of a fusion protein
                       containing the DNA binding domain;
                 (iv) a transcription termination signal
                       operably linked to the first nucleotide
                       sequence;
20   (v)   a first means for replicating in the
                       cells of said yeast strains in (a) and (b);
                       and
           (d) a second vector comprising:
                 (i)   a second promoter;
25   (ii) a nucleotide sequence encoding an
                       activation domain of a transcriptional
                       activator, operably linked to the second
                       promoter;
                 (iii) means for inserting a DNA sequence
30   encoding a protein into the vector in such a
                       manner that the protein is capable of being
                       expressed as part of a fusion protein
                       containing the activation domain of a
                       transcriptional activator;
35   (iv) a transcription termination signal
                       operably linked to the second nucleotide
                       sequence; and

(v)   a second means for replicating in the
purified cells of said yeast strains in (a)
and (b).

5   90.   The method according to claim 50 in which said
recombinant nucleic acids encoding said one or more candidate
molecules each comprise the following operably linked
components:
            (a) an ADC1 promoter;
10   (b) a nucleotide sequence encoding a candidate
       molecule fused to a nuclear localization signal; and
            (c) an ADC1 transcription termination signal.

       91.   A purified expression vector comprising the
15 following components:
            (a) a promoter active in yeast;
            (b) a first nucleotide sequence encoding a peptide
       of 20 or fewer amino acids fused to a nuclear
       localization signal, said first nucleotide sequence
20   being operably linked to the promoter;
            (c) a transcription termination signal active in
       yeast, operably linked to said first nucleotide
       sequence;
            (d) means for replicating in a yeast cell;
25   (e) means for replicating in E. coli;
            (f) a second nucleotide sequence encoding a
       selectable marker for selection in a yeast cell,
       operably linked to a transcriptional promoter and
       transcription termination signal active in yeast; and
30   (g) a third nucleotide sequence encoding a
       selectable marker for selection in E. coli, operably
       linked to a transcriptional promoter and transcription
       termination signal active in E. coli.

35   92.   The method according to claim 50 in which diploid
       yeast cells have a mutation in at least one nucleic acid
       coding for a cell wall component thereby having a modified

cell wall that is more permeable to exogenous molecules than
is a wild-type cell wall.

93.   The method according to claim 50 in which diploid
5 yeast cells have a mutation in at least one nucleic acid
coding for a cell wall component thereby having a modified
cell wall that is more permeable to exogenous molecules than
is a wild-type cell wall.

10   94.   The method according to claim 54 in which said
environment of incubating step (h) contains
3-amino-1,2,4-triazole.

95.   The method according to claim 86 in which said
15 third environment contains 3-amino-1,2,4-triazole.

96.   A method of detecting an inhibitor of a protein-
protein interaction comprising
          (a) incubating a population of cells, said
20   population comprising cells recombinantly expressing a
          pair of interacting proteins, said pair consisting of a
          first fusion protein and a second fusion protein, in the
          presence of one or more candidate molecules among which
          it is desired to identify an inhibitor of the
25   interaction between said first fusion protein and said
          second fusion protein, each said first fusion protein
          comprising a first protein sequence and a DNA binding
          domain; each said second fusion protein comprising a
          second protein sequence and a transcriptional activation
30   domain of a transcriptional activator; and in which the
          cells contain a first nucleotide sequence operably
          linked to a promoter driven by one or more DNA binding
          sites recognized by said DNA binding domain such that an
          interaction of said first fusion protein with said
35   second fusion protein results in increased transcription
          of said first nucleotide sequence, said incubating being
          in an environment in which substantial death of said

- 357 -

cells occurs (i) when said increased transcription
occurs of said first nucleotide sequence or (ii) if said
cells lack a recombinant nucleic acid encoding said
first fusion protein or a recombinant nucleic acid
5   encoding said second fusion protein; and
           (b)   detecting those cells that survive said
incubating step, thereby detecting the presence of an
inhibitor of said interaction in said cells.

10   97.   The method according to claim 96 in which said
population of cells comprises a plurality of cells, each cell
within said plurality recombinantly expressing a different
said pair of interacting proteins.

15   98.   The method according to claim 96 in which the cells
are yeast cells.

       99.   The method according to claim 97 in which the cells
are yeast cells.
20

       100.   The method according to claim 98 or 99 in which
the first nucleotide sequence is functional *URA3* coding
sequences, and said environment contains 5-fluoroorotic acid.

25   101.   The method according to claim 96 in which said one
or more candidate molecules are provided to said cells by
introducing into said cells one or more recombinant nucleic
acids encoding said one or more candidate molecules, such
that said one or more candidate molecules are expressed
30 within said cells.

       102.   The method according to claim 96 in which said
environment contains said one or more candidate molecules.

35   103.   The method according to claim 98 in which the
cells are haploid yeast cells.

104.   The method according to claim 98 in which the cells are diploid yeast cells.

105.   The method according to claim 17 in which cells are yeast cells, and in which the first protein and the second protein are first and second fusion proteins, respectively, between which an interaction is detected according to the method of claim 54.

106.   The method according to claim 97 or 99 in which the plurality of cells consists of at least 10 cells.

107.   The method according to claim 97 or 99 in which the plurality of cells consists of at least 100 cells.

108.   The method according to claim 97 or 99 in which the plurality of cells consists of at least 1000 cells.

109.   The method according to claim 96 in which said one or more candidate molecules are compounds synthesized by proteins encoded by recombinant DNA that has been introduced into said cells.

110.   The method according to claim 17 in which said population of cells comprises a plurality of cells, each cell within said plurality recombinantly expressing a different said pair of interacting proteins.

111.   The method according to claim 110 in which the plurality of cells consists of at least 10 cells.

112.   The method according to claim 110 in which the plurality of cells consists of at least 100 cells.

113.   The method according to claim 110 in which the plurality of cells consists of at least 1000 cells.

114. A method of detecting one or more protein-protein
interactions comprising:

(a) recombinantly expressing in a population of
host cells a first population of first fusion proteins,

5   wherein each said first fusion protein comprises a first
protein sequence and a DNA binding domain of a
transcriptional activator, in which the DNA binding
domain is the same in each said first fusion protein,
and wherein said host cells contain at least one

10   nucleotide sequence operably linked to a promoter driven
by one or more DNA binding sites recognized by said DNA
binding domain, such that an interaction of a first
fusion protein with a second fusion protein in a host
cell, said second fusion protein comprising a

15   transcriptional activation domain, results in increased
transcription in said host cell of said at least one
nucleotide sequence;

(b) negatively selecting said population of host
cells to reduce the number of said host cells expressing

20   said first population of first fusion proteins in which
said increased transcription of said at least one
nucleotide sequence occurs in the absence of said second
fusion protein to less than 5×10^{-6} of the total number
of host cells;

25   (c) recombinantly expressing in said negatively-
selected population of host cells a second population of
second fusion proteins, wherein each said second fusion
protein comprises a second protein sequence and an
activation domain of a transcriptional activator, in

30   which the activation domain is the same in each said
second fusion protein, such that a first fusion protein
is co-expressed with a second fusion protein in said
host cells; and

(d) detecting increased transcription of said at

35   least one nucleotide sequence; thereby detecting an
interaction between a first fusion protein and a second
fusion protein.

- 360 -

115. The method according to 114 wherein said step of negative selecting reduces the number of said host cells in which said increased transcription of said at least one nucleotide sequence occurs in the absence of said second
5 fusion protein to less than 1×10^{-6} of the total number of host cells.

116. The method according to claim 114 wherein said increased transcription of said at least one nucleotide
10 sequence renders said host cell sensitive to toxic effects of a chemical agent which is otherwise non-toxic in the absence of said increased transcription, and wherein said step of negative selecting comprises:
            (a)   a first growing of said population of host
15   cells in a first environment containing said chemical
        agent; and
            (b)   a second growing of a plurality of cells in a
        second environment containing said chemical agent,
        wherein said plurality of cells comprises growing cells
20   from said first growing.

117. The method according to claim 116 wherein said first growing is on a first solid medium containing said chemical agent, wherein said second growing is on a second
25 solid medium containing said chemical agent, and further comprising, between said first growing step and said second growing step, a step of physically transferring cells from colonies of growing cells from said first environment to said second environment.
30

118.   The method according to claim 117 where said physically transferring is by replica plating cells from said first solid medium to said second solid medium.

35   119. The method according to claim 116 further comprising, after said second growing step, a third growing of cells surviving said second growing in a third environment

- 361 -

containing said chemical agent such that said third growing independently negatively selects host cells surviving said first and second growings and in which said increased transcription of said at least one nucleotide sequence occurs
5 in the absence of said second fusion proteins.

120. The method according to claim 116 wherein said increased transcription of said at least one nucleotide sequence in a host cell confers a URA+ phenotype on said host
10 cell, and wherein said chemical agent is 5-fluoroorotic acid.

121. The method according to claim 114 wherein said at least one nucleotide sequence comprises at least a first nucleotide sequence and a second nucleotide sequence, wherein
15 increased transcription of said first nucleotide sequence renders said host cell sensitive to lethal effects of a first chemical agent which is otherwise non-toxic in the absence of said increased transcription of said first nucleotide sequence, wherein increased transcription of said second
20 nucleotide sequence renders said host cell sensitive to lethal effects of a second chemical agent which is otherwise non-toxic in the absence of said increased transcription of said second nucleotide sequence, and wherein said step of negative selecting comprises:
25   (a)   a first growing of said population of host
         cells in a first environment containing said first
         chemical agent such that those host cells are negatively
         selected in which said increased transcription of said
         first nucleotide sequence occurs in the absence of said
30   second fusion proteins; and
             (b)   a second growing of cells surviving said first
         growing in a second environment containing said second
         chemical agent such that said second growing
         independently negatively selects host cells in which
35   said increased transcription of said second nucleotide
         sequence occurs in the absence of said second fusion
         proteins.

- 362 -

122. The method according to claim 121 wherein said first and second environment are a single environment containing both said first and said second chemical agents.

5   123. The method according to claim 114 wherein before said recombinantly expressing step (c) the following steps are carried out:

          (a) recovering cells expressing a chosen one of said first fusion proteins from said population of host
10   cells recombinantly expressing a population of first fusion proteins;

          (b) recombinantly expressing in said cells expressing said chosen first fusion protein one or more of said second fusion proteins, such that said chosen
15   first fusion protein is co-expressed with a second fusion protein in said recovered cells; and

          (c) detecting the rate of increased transcription of said at least one nucleotide sequence as the fraction of second fusion proteins which cause said increased
20   transcription when co-expressed with said chosen first fusion protein.

124. The method according to claim 123 wherein there are at least 50,000 of said second fusion proteins.
25

125. The method according to claim 114 wherein before said recombinantly expressing step (c) the following steps are carried out:

          (a) recovering cells expressing one or more of said
30   first fusion proteins from said population of host cells recombinantly expressing a population of first fusion proteins;

          (b) recombinantly expressing in said cells expressing one or more of said first fusion proteins a
35   chosen one of said second fusion proteins, such that said one or more first fusion proteins are co-expressed

with said chosen second fusion protein in said recovered
cells; and

    (c) detecting the rate of increased transcription
of said at least one nucleotide sequence as the fraction
5   of first fusion proteins which cause said increased
transcription when co-expressed with said chosen second
fusion protein.

    126. The method according to claim 125 wherein there are
10 at least 50,000 of said first fusion proteins.

    127. The method according to claim 123 or 125 wherein
said first or second fusion protein is not subjected to the
recombinantly expressing step (c) or the detection step (d)
15 of claim 114 if said rate of increased transcription is
greater than 10^{-3}.

    128. The method according to claim 114 comprising, after
said detecting step (d), the steps of:
20   (e) further selecting said population of host
cells for absence of a second fusion protein; and
           (f) detecting in said further selected cells
said increased transcription of said at least one
nucleotide sequence;
25   whereby cells are detected in which increased
transcription of said at least one nucleotide sequence occurs
in the absence of said second fusion protein.

    129. The method according to claim 128 comprising, after
30 said detecting step (f), a second step of negatively
selecting said cells in which increased transcription of said
at least one nucleotide sequence occurs in the absence of
said second fusion protein.

35   130. The method according to claim 128 wherein said
second population of second fusion proteins is expressed from
a plasmid expression vector that expresses a selectable

marker, and wherein said step (e) of further selecting
comprises incubating said population of host cells in an
environment in which substantial death of host cells occurs
in the absence of expression of the selectable marker.

5

     131. The method according to claim 114 comprising, after
said detecting step (d), the steps of:

          (e) recombinantly expressing in a second
population of host cells those first fusion

10

proteins that are recombinantly expressed in said
host cells in which said interaction is detected in
step (d);

          (f) recombinantly expressing in said second
population of host cells those second fusion

15

proteins that are recombinantly expressed in said
host cells in which said interaction is detected in
step (d), such that a first fusion protein is co-
expressed with a second fusion protein in said
second population of host cells; and

20

          (g) selecting said host cells co-expressing a
first and a second fusion protein for increased
transcription of said at least one nucleotide
sequence.

25

     132. The method according to claim 131 further
comprising, after said selecting step (g), a step of
negatively selecting among said selected host cells those
expressing a first fusion protein that is co-expressed along
with a majority of other second fusion proteins in said cells

30

selected in step (g).

     133. The method according to claim 131 further
comprising, after said selecting step (g), a step of
negatively selecting among said selected host cells those

35

expressing a second fusion protein that is co-expressed along
with a majority of other first fusion proteins in said cells
selected in step (g).

134.   The method according to claim 131 wherein said
first population of first fusion proteins is expressed from a
first plasmid expression vector, and wherein said expressing
step (e) comprises the steps of:

5   (a) rescuing said first plasmid expression vectors
        from said host cells in which said interaction between a
        first fusion protein and a second fusion protein is
        detected in step (d); and
            (b) transforming said second population of host
10   cells with said rescued first plasmid expression
        vectors.

135.   The method according to claim 134 wherein said
second population of second fusion proteins is expressed from
15 a second plasmid expression vector, and wherein said
expressing step (f) comprises the steps of:

            (a) rescuing said second plasmid expression vectors
        from said host cells in which said interaction between a
        first fusion protein and a second fusion protein is
20   detected in step (d); and
            (b) transforming said second population of host
        cells with said rescued second plasmid expression
        vectors.

25   136.   A method of detecting one or more protein-protein
interactions comprising:

            (a)   recombinantly expressing in a first population
        of host cells a first population of first fusion
        proteins, wherein each said first fusion protein
30   comprises a first protein sequence and a DNA binding
        domain of a transcriptional activator, in which the DNA
        binding domain is the same in each said first fusion
        protein, and wherein said host cells contain at least
        one nucleotide sequence operably linked to a promoter
35   driven by one or more DNA binding sites recognized by
        said DNA binding domain, such that an interaction of a
        first fusion protein with a second fusion protein in a

host cell, said second fusion protein comprising a
transcriptional activation domain, results in increased
transcription in said host cell of said at least one
nucleotide sequence;

5   (b)   recombinantly expressing in said first
population of host cells a second population of second
fusion proteins, wherein each said second fusion protein
comprises a second protein sequence and an activation
domain of a transcriptional activator, in which the
10   activation domain is the same in each said second fusion
protein, such that a first fusion protein is
co-expressed with a second fusion protein in said first
population of host cells;

           (c)   selecting said host cells in said first
15   population of host cells for increased transcription of
said at least one nucleotide sequence;

           (d)   recombinantly expressing in a second
population of said host cells a third population of
third fusion proteins, wherein each said third fusion
20   protein comprises said second protein sequence and said
DNA binding domain of a transcriptional activator, in
which the DNA binding domain is the same in each said
third fusion protein;

           (e)   recombinantly expressing in said second
25   population of host cells a fourth population of fourth
fusion proteins, wherein each said fourth fusion protein
comprises said first protein sequence and an activation
domain of said transcriptional activator, in which the
activation domain is the same in each said fourth fusion
30   protein, such that a third fusion protein is
co-expressed with a fourth fusion protein in said second
population of host cells;

           (f)   selecting said host cells in said second
population of host cells for increased transcription of
35   said at least one nucleotide sequence; thereby detecting
an interaction between a first fusion protein and a
second fusion protein.

137. The method according to claim 136 comprising, after step (f), the following steps:

      (g)   determining the identities of first pairs of said first and said second protein sequences in said

5   host cells selected in step (c);

      (h)   determining the identities of second pairs of said first and said second protein sequences in said host cells selected in step (f); and

      (i)   selecting as bi-directional interacting pairs of

10 said first and said second protein sequences those pairs of said first and said second protein sequences that are found among both said first pairs and said second pairs of said first and said second protein sequences.

15   138.   A method of detecting one or more protein-protein interactions comprising:

      (a)   recombinantly expressing in a first population of yeast cells of a first mating type a first population of first fusion proteins, wherein each first fusion

20   protein comprises a first protein sequence and a DNA binding domain of a transcriptional activator, in which the DNA binding domain is the same in each said first fusion protein, and wherein said first population of yeast cells contains at least one nucleotide sequence

25   operably linked to a promoter driven by one or more DNA binding sites recognized by said DNA binding domain, such that an interaction of a first fusion protein with a second fusion protein in a yeast cell, said second fusion protein comprising a transcriptional activation

30   domain, results in increased transcription in said yeast cell of said at least one nucleotide sequence;

      (b)   negatively selecting said first population of yeast cell to reduce the numbers of those yeast cells expressing said first population of first fusion

35   proteins in which said increased transcription of said at least one nucleotide sequence occurs in the absence of said second fusion protein;

(c)   recombinantly expressing in a second population of yeast cells of a second mating type, different from said first mating type, a second population of said second fusion proteins, wherein each second fusion protein comprises a second protein sequence and an activation domain of a transcriptional activator, in which the activation domain is the same in each said second fusion protein;

(d)   mating said first population of yeast cells with said second population of yeast cells to form a population of diploid yeast cells, wherein said mating occurs on a solid support with a cell density of greater than 5 x 10^4 cells per square millimeter of solid support, such that a first fusion protein is co-expressed with a second fusion protein in said diploid cells; and

(e)   selecting said host cells co-expressing a first and a second fusion protein for increased transcription of said at least one nucleotide sequence; thereby detecting an interaction between a first fusion protein and a second fusion protein.

139. The method according to claim 138 wherein said cell density is between 1.5 x 10^5 cells and 4 x 10^5 cells per square millimeter.

140. The method according to claim 138 wherein said solid support is a filter having a pore size sufficiently small to retain said yeast cells.

141.   A method of detecting and recording one or more protein-protein interactions comprising:

(a)   recombinantly expressing within a population of host cells

(i)   a first population of first fusion proteins, each said first fusion protein comprising a first protein sequence and a DNA binding domain,

in which the DNA binding domain is the same in each
said first fusion protein, and

(ii) a second population of second fusion
proteins, each said second fusion protein
comprising a second protein sequence and a
transcriptional activation domain of a
transcriptional activator, in which the
transcriptional activation domain is the same in
each said second fusion protein, such that a first
fusion protein is co-expressed with a second fusion
protein in host cells, and wherein said host cells
contain at least one nucleotide sequence operably
linked to a promoter driven by one or more DNA
binding sites recognized by said DNA binding domain
such that interaction of a first fusion protein
with a second fusion protein results in activation
of transcription of said at least one nucleotide
sequence by said regulatory domain;

(b) selecting as positive those host cells that
co-express a first fusion protein and a second fusion
protein and that have increased transcription of said at
least one nucleotide sequence; and

(c) updating a first computer-implemented data-
store with (i) information in digital form
characterizing a plurality of said selected positive
host cells, and with (ii) information in digital form
characterizing said first protein sequences and said
second protein sequences in a plurality of said selected
positive cells.

142. The method according to claim 141 wherein steps (b)
and (c) are repeated for a plurality of selected positive
host cells having different said first and said second
protein sequences.

143. The method according to claim 141 wherein said
information characterizing said first and said second protein

sequences comprises QEA™ or SEQ-QEA™ signals derived from
nucleic acids encoding first and said second protein
sequences.

5   144. The method according to claim 141 wherein said
information characterizing said first and said second protein
sequences comprises first and second nucleotide sequences of
first and second nucleic acids encoding said first and said
second protein sequences.
10

        145. The method according to claim 144 further
comprising, after said step of updating, the steps of:
        (d)   searching a nucleotide sequence database
        comprising nucleic acid coding sequences for one or more
15   nucleic acid coding sequences that are homologous to, or
        the absence of any nucleic acid coding sequences that
        are homologous to, said first or said second nucleotide
        sequences;
        (e) retrieving sequence-identifying information in
20   digital form for each homologous nucleic acid coding
        sequence, said sequence-identifying information
        comprising (i) a characterization of the identity of the
        homologous nucleic acid coding sequence, (ii) the degree
        of homology of the homologous nucleic acid coding
25   sequence with said first or said second nucleotide
        sequence, and (iii) the location of said first or said
        second nucleotide sequence in the homologous nucleic
        acid coding sequence; and
        (f) updating said first computer-implemented data-
30   store with said retrieved sequence-identifying
        information for each homologous nucleic acid coding
        sequence for said first and said second protein
        sequences.

35   146. The method according to claim 145 further
comprising, after said step of updating, the steps of:

- 371 -

(g) choosing as a first gene to represent the origin of said first protein sequence either (i) one of said retrieved homologous nucleic acid coding sequences, or (ii) said first nucleotide sequence and an indication

5   that said first nucleotide sequence is a new sequence;

(h) choosing as a second gene to represent the origin of said second protein sequence either (i) one of said retrieved homologous nucleic acid coding sequences, or (ii) said second nucleotide sequence and an

10   indication that said second nucleotide sequence is a new sequence;

(i) updating a second computer-implemented data-store with information in digital form comprising a representation (A) that said first and said second genes

15   code for proteins that participate in a protein-protein interaction and (B) that said selected positive cells evidence said protein-protein interaction; and

(j) updating said second computer-implemented data-store with information in digital form comprising a

20   representation (i) that said selected positive host cells co-express protein sequences whose origin is said first and said second genes, and (ii) a representation of the locations of said first and said second nucleotide sequences on the nucleic acid coding

25   sequences of said first and said second genes, respectively.

147. The method according to claim 146 wherein said first computer-implemented data-store and said second

30 computer-implemented data-store are the same computer-implemented data-store.

148. The method according to claim 146 further comprising a step of updating said second computer-

35 implemented data-store with information in digital form representing the results of confirmation tests that confirm that said increased transcription of said selected positive

- 372 -

host cells accurately reflects interaction of said first
fusion protein and said second fusion protein.

149. The method according to claim 146 wherein said
5 steps (g) and (h) of choosing comprise applying at least one
of the following rules:

       (i) eliminating from consideration for choice
homologous nucleic acid coding sequences with a
species origin different than the species origin of
10   said protein sequences,

       (ii) eliminating from consideration for
choice homologous nucleic acid coding sequences
that are anti-sense, or

       (iii) eliminating from consideration for
15   choice all pairs of homologous nucleic acid coding
sequences whose coded proteins have different
general cellular functions.

150. The method according to claim 149 further
20 comprising choosing the homologous nucleic acid coding
sequences having the greatest degree of homology to said
first or second nucleotide sequence.

151. The method according to claim 146 wherein said
25 steps (g) and (h) of choosing are performed by a rule-based
program.

152. The method according to claim 146 further
comprising the steps of:
30   (k) retrieving from said second computer-
implemented data-store information in digital form
representing said protein-protein interactions of a
selected subset of said first or said second genes;

      (l) determining one or more connected components of
35   a graph representation of said retrieved protein-protein
interactions for said selected subset of said first or
said second genes; and

(m) outputting a representation of said connected
components;

whereby putative protein interaction pathways are
determined.

5

153. The method according to claim 146 further
comprising the steps of:

(k) retrieving from said second computer-
implemented data-store information representing, for a
10   selected protein-protein interaction, said locations of
said first and said second nucleotide sequences on the
sequences of said first and said second genes for those
selected positive host cells evidencing said protein-
protein interaction;

15   (l) intersecting said retrieved locations in said
first gene and in said second gene in order to find
domains of intersection in said first gene and in said
second gene; and

(m) outputting a representation of said retrieved
20   locations and said domains of intersection;
whereby the physical domain of said protein-protein
interaction is approximated.

154. The method according to claim 153, wherein said
25 second computer-implemented data-store comprises information
representing that a first protein originating from a first
gene interacts with a second protein originating from a
second gene and that a third protein originating from a third
gene also interacts with said second protein, and wherein
30 said domain of intersection on said second protein determined
according to claim 153, wherein said selected protein-protein
interaction is the interaction of said first and said second
proteins, overlaps with said domain of intersection on said
second protein determined according to claim 152, wherein
35 said selected protein-protein interaction is the interaction
of said third and said second proteins, and wherein said
method further comprises searching for homologies between

- 374 -

said domain of interaction on said first protein and said
domain of interaction on said third protein.

155. A computer-implemented method for storing and
5 analyzing at least one pair-wise interaction between protein
sequences coded for by nucleic acids originating from
biological samples, said method comprising:
          (a) searching a nucleotide sequence database,
       comprising nucleic acid coding sequences from biological
10   samples, for
                   (i) one or more nucleic acid coding sequences
               that are homologous to, or the absence of any
               nucleic acid coding sequences that are homologous
               to, a first or a second nucleotide sequence,
15   wherein said pair-wise interaction comprises an
               interaction between a first and a second protein
               sequence, and wherein said first and said second
               nucleotide sequences are sequences of said nucleic
               acids coding for said first and said second protein
20   sequence, respectively, and
                   (ii) retrieving sequence-identifying
               information in digital form for each homologous
               nucleic acid coding sequence, said sequence-
               identifying information comprising (i) the identity
25   of a homologous nucleic acid coding sequence and
               (ii) the location of said first or second
               nucleotide sequence on the homologous nucleic acid
               coding sequence;
          (b) choosing as a first gene to represent the
30   origin of said first protein sequence either (i) one of
       said retrieved homologous nucleic acid coding sequences,
       or (ii) said first nucleotide sequence and an indication
       that said first nucleotide sequence is a new sequence;
          (c) choosing as a second gene to represent the
35   origin of said second protein sequence either (i) one of
       said retrieved homologous nucleic acid coding sequences,
       or (ii) said second nucleotide sequence and an

indication that said second nucleotide sequence is a new sequence; and

       (d) updating a first computer-implemented data-store with information in digital form comprising a

5   representation (i) that said first and second genes code for proteins that participate in a protein-protein interaction, and (ii) that said pair-wise interaction of protein sequences evidences said protein-protein interaction.

10

      156. The method according to claim 155 wherein steps (a)-(d) are repeated for a plurality of pair-wise interactions between protein sequences.

15   157. The method according to claim 156 wherein a plurality of said pair-wise interactions evidence one said protein-protein interaction.

      158. The method according to claim 155 further
20 comprising updating said first computer-implemented data-store with information in digital form comprising a representation of the identities of said first and said second genes.

25   159. The method according to claim 155 further comprising, before said step of searching, a step of determining a pair-wise interaction between two protein sequences by a method comprising the reconstitution of a transcriptional activator in a host cell due to interaction
30 of said first and said second protein sequences in said host cell.

      160. The method according to claim 155 further comprising the steps of:
35   (a) retrieving from said first computer-implemented data-store information in digital form representing said

protein-protein interactions of a selected subset of
said first or said second genes;

     (b) determining one or more connected components of
a graph representation of said retrieved protein-protein
5   interaction for said selected subset of said first or
said second genes; and

     (c) outputting a representation of said connected
components;

whereby putative protein interaction pathways are
10 determined.

161. The method according to claim 155 further
comprising, after said choosing steps (b) and (c), a step of
updating said first computer-implemented data-store with
15 information in digital form comprising a representation (i)
that said pair-wise interaction of protein sequences
comprises protein sequences whose origin is said first gene
and said second gene, and (ii) of said locations of said
first and said second nucleotide sequences in the sequences
20 of said first and said second genes.

162. The method according to claim 161 further
comprising the steps of:

     (e) retrieving from said first computer-implemented
25   data-store information representing, for a selected
protein-protein interaction, said locations of said
first and said second nucleotide sequences on the
sequences of said first and said second genes for pair-
wise interactions evidencing said protein-protein
30   interaction;

     (f) intersecting said retrieved locations in said
first gene and in said second gene in order to find
domains of intersection in said first gene and in said
second gene; and

35   (g) outputting a representation of said retrieved
locations and said domains of intersection;

- 377 -

whereby the physical domain of said protein-protein
interaction is approximated.

163. The method according to claim 162, wherein said
5 first computer-implemented data-store comprises information
representing that a first protein originating from a first
gene interacts with a second protein originating from a
second gene and that a third protein originating from a third
gene also interacts with said second protein, and wherein
10 said domain of intersection on said second protein determined
according to claim 161, wherein said selected protein-protein
interaction is the interaction of said first and said second
proteins, overlaps with said domain of intersection on said
second protein determined according to claim 161, wherein
15 said selected protein-protein interaction is the interaction
of said third and said second proteins, and further .   .
comprising searching for homologies between said domain of
interaction on said first protein and said domain of
interaction on said third protein.
20

164. The method according to claim 161 further
comprising updating said first computer-implemented data-
store with information in digital form comprising an
identification of said pair-wise interaction.
25

165. The method according to claim 155 wherein said
steps (b) and (c) of choosing comprise applying at least one
of the following rules:
                    (i)   eliminating from consideration for choice
30   homologous nucleic acid coding sequences with a
            species origin different than the species origin of
            said protein sequences,
                    (ii) eliminating from consideration for
            choice homologous nucleic acid coding sequences
35   that are anti-sense, or
                    (iii) eliminating from consideration for
            choice all pairs of homologous nucleic acid coding

sequences whose coded proteins have different
general cellular functions.

166. The method according to claim 165, wherein said
sequence-identifying information further comprises a degree
of homology of a homologous nucleic acid coding sequence with
said first or said second nucleotide sequence, and wherein
said method further comprises choosing as a first or a second
gene the homologous nucleic acid coding sequences having the
greatest degree of homology to said first or second
nucleotide sequence.

167. The method according to claim 155 wherein said
steps (b) and (c) of choosing are performed by a rule-based
program.

168. The method according to claim 155 wherein said
steps (b) and (c) of choosing comprise at least one user
input.

169. The method according to claim 155 wherein said
sequence-identifying information further comprises a degree
of homology of the homologous nucleic acid coding sequence
with said first or said second nucleotide sequence, and where
said method further comprises updating a second computer-
implemented data-store with said retrieved sequence-
identifying information for each homologous nucleic acid
coding sequence for said first or said second protein
sequences.

170. A computer-readable medium containing instructions
for causing one or more computers to function according to
the method of claim 155.

171. A computer system for storing and processing data
related to at least one pair-wise interaction between protein
sequences encoded by nucleic acids originating from

biological samples, said computer system comprising at least
one computer memory, said computer memory comprising data
structures for information in digital form representing

(a) an identity of a selected first gene, wherein
said pair-wise interaction comprises an interaction
between a first and a second protein sequence, and
wherein said first gene comprises a first coding
sequence homologous to a first nucleotide sequence of a
nucleic acid coding for said first protein sequence;

(b) an identity of a selected second gene, wherein
said second gene comprises a second coding sequence
homologous to a second nucleotide sequence of a nucleic
acid coding for said second protein sequence of said
pair-wise interaction;

(c) an indication that said first and said second
genes code for proteins involved·in a protein-protein
interaction;

(d) an indication that said pair-wise interaction
evidences said protein-protein interaction;

(e) a first location of said first nucleotide
sequence on the coding sequence of said first gene; and

(f) a second location of said second nucleotide
sequence on the coding sequence of said second gene.

172. The computer system according to claim 171 wherein
said data structures are in a relational database format.

173. The computer system according to claim 171 wherein
said computer memory further comprises data structures for a
plurality of said protein-protein interactions and for a
plurality of said pair-wise interactions between protein
sequences.

174. The computer system according to claim 173 wherein
said computer memory further comprises data structures for
information in digital form representing a plurality of said

pair-wise interactions which evidence one of said protein-protein interactions.

175. The computer system according to claim 171 further comprising at least one computer for processing one or more computer programs, said computer programs for:

(a) searching a nucleotide sequence database comprising nucleic acid coding sequences from biological samples, for one or more nucleic acid coding sequences that are homologous to, or the absence of any nucleic acid coding sequences that are homologous to, said first or said second nucleotide sequences;

(b) retrieving sequence-identifying information in digital form for each homologous nucleic acid coding sequence, said sequence-identifying information comprising (i) the identity of a homologous nucleic acid coding sequence and (ii) the location of said first or said second nucleotide sequence on the homologous nucleic acid coding sequence;

(c) choosing as a first gene to represent the origin of said first protein sequence either (i) one of said retrieved homologous nucleic acid coding sequences, or (ii) said first nucleotide sequence and an indication that said first nucleotide sequence is a new sequence;

(d) choosing as a second gene to represent the origin of said second protein sequence either (i) one of said retrieved homologous nucleic acid coding sequences, or (ii) said second nucleotide sequence and an indication that said second nucleotide sequence is a new sequence; and

(e) updating in said computer memory said data structures with information in digital form comprising a representation of

(i) the identity of said first chosen gene,

(ii) the identity of said second chosen gene,

(iii) an indication that said first and said second gene encode proteins involved in a protein-protein interaction,

(iv) an indication that said pair-wise interaction evidences said protein-protein interaction,

(v) said first location of said first nucleotide sequence on the coding sequence of said first gene, and

(vi) said second location of said second nucleotide sequence on the coding sequence of said second gene.

176. A computer readable memory for storing data related to at least one pair-wise interaction between protein sequences encoded by nucleic acids originating from biological samples, said computer readable memory comprising data structures for receiving information in digital form representing

(a) an identity of a selected first gene, wherein said pair-wise interaction comprises an interaction between a first and a second protein sequence, and wherein said first gene comprises a first coding sequence homologous to a first nucleotide sequence of a nucleic acid encoding said first protein sequence;

(b) an identity of a selected second gene, wherein said second gene comprises a second coding sequence homologous to a second nucleotide sequence of a nucleic acid encoding a second protein sequence of said pair-wise interaction between protein sequences;

(c) an indication that said first and said second genes encode proteins involved in a protein-protein interaction;

(d) an indication that said pair-wise interaction evidences said protein-protein interaction;

(e) a first location of said first nucleotide sequence on the coding sequence of said first gene; and

(f)   a second location of said second nucleotide
sequence on the coding sequence of said second gene.

177. The computer readable memory according to claim 176
5 wherein said data structures are in a relational database
format.

178. The computer readable memory according to claim 176
wherein said computer memory further comprises data
10 structures for a plurality of said protein-protein
interactions and for a plurality of said pair-wise
interactions between protein sequences.

179. The computer readable memory according to claim 178
15 wherein said computer readable memory further comprises data
structures for receiving information in digital form
representing a plurality of said pair-wise interactions which
evidence one of said protein-protein interactions.

20   180.   A method of detecting one or more protein-protein
interactions comprising:
        (a)   recombinantly expressing in a population of
        host cells a first population of first fusion proteins,
        wherein each said first fusion protein comprises a first
25   protein sequence and a DNA binding domain of a
        transcriptional activator, in which the DNA binding
        domain is the same in each said first fusion protein,
        and wherein said host cells contain at least one
        nucleotide sequence operably linked to a promoter driven
30   by one or more DNA binding sites recognized by said DNA
        binding domain, such that an interaction of a first
        fusion protein with a second fusion protein, said second
        fusion protein comprising a transcriptional activation
        domain, results in increased transcription in said host
35   cell of said at least one nucleotide sequence;
        (b)   negatively selecting said population of host
        cells to reduce the fraction of said host cells

expressing said first population of first fusion
proteins in which said increased transcription of said
at least one nucleotide sequence occurs in the absence
of said second fusion protein, said negative selecting
5   being by a method comprising transferring one or more
times growing host cells between environments in which
substantial cell death occurs upon activation of said at
least one nucleotide sequence;
     (c)   recombinantly expressing in said negatively-
10   selected population of host cells a second population of
second fusion proteins, wherein each said second fusion
protein comprises a second protein sequence and an
activation domain of a transcriptional activator, in
which the activation domain is the same in each said
15   second fusion protein, such that a first fusion protein
is co-expressed with a second fusion protein in said
host cells; and
     (d)   selecting said host cells co-expressing a
first and a second fusion protein for increased
20   transcription of said at least one nucleotide sequence;
thereby detecting an interaction between a first fusion
protein and a second fusion protein.

181. The method according to claim 180 wherein after
25 said negative selective step said fraction is less than or
equal to 5 x 10^{-6}.

182. The method according to claim 180 wherein after
said negative selective step said fraction is less than or
30 equal to 1 x 10^{-6}.

183. The method according to claim 180 wherein increased
transcription of said at least one nucleotide sequence
renders said host cell sensitive to toxic effects of a
35 chemical agent which is otherwise non-toxic in the absence of
said increased transcription, said toxic effects comprising

substantial cell death, and wherein said negative selecting
comprises the steps of:

     (a) a first growing of said population of host
cells in a first environment containing said chemical
5 agent; and

     (b) a second growing of a plurality of cells in a
second environment containing said chemical agent,
wherein said plurality of cells comprises growing cells
transferred from said first growing.

10

    184. The method according to claim 183 wherein said
first growing is on a first solid medium containing said
chemical agent, wherein said second growing is on a second
solid medium containing said chemical agent, and further
15 comprising, between said first growing step and said second
growing step, a step of physically transferring cells from
colonies of growing cells from said first environment to said
second environment.

20   185. The method according to claim 184 wherein said
physically transferring is by replica plating cells from said
first solid medium to said second solid medium.

    186. The method according to claim 183 which further
25 comprises, after said second growing, a third growing of a
second plurality of cells in a third environment containing
said chemical agent, wherein said second plurality of cells
comprises growing cells transferred from said second growing.

30   187. The method according to claim 180 wherein said at
least one nucleotide sequence comprises at least a first
nucleotide sequence and a second nucleotide sequence, wherein
increased transcription of said first nucleotide sequence
renders said host cell sensitive to toxic effects of a first
35 chemical agent, which is otherwise non-toxic in the absence
of said increased transcription of said first nucleotide
sequence, wherein increased transcription of said second

nucleotide sequence renders said host cell sensitive to toxic
effects of a second chemical agent, which is otherwise non-
toxic in the absence of said increased transcription of said
second nucleotide sequence, wherein said toxic effects of
5 said first and said second chemical agent comprise
substantial cell death, and wherein said negative selecting
comprises the steps of:

(a) a first growing of said population of host
cells in a first environment containing said first
10   chemical agent; and

(b) a second growing of a plurality of cells in a
second environment containing said second chemical
agent, wherein said plurality of cells comprises growing
cells transferred from said first growing.
15

188. The method according to claim 100 in which said one
or more candidate molecules are provided to said cells by
introducing into said cells one or more recombinant nucleic
acids encoding said one or more candidate molecules, such
20 that said one or more candidate molecules are expressed
within said cells.

189. The method according to claim 100 in which said
environment contains said one or more candidate molecules.
25

190. The method according to claim 100 in which said
one or more candidate molecules are compounds synthesized by
proteins encoded by recombinant DNA that has been introduced
into said cells.
30

35

CONSTRUCT DNA-BINDING & ACTIVATION DOMAIN FUSION LIBRARIES FROM
DIFFERENT POPULATIONS (TISSUE/DEVELOPMENT STAGE/DISEASE STAGE)

↓

PERFORM YEAST MATING SCREENS TO IDENTIFY INTERACTING PAIRS OF PROTEINS
IN EACH POPULATION - ARRIVE AT "INTERACTIVE POPULATIONS"

↓

CHARACTERIZE INTERACTIVE POPULATIONS BY A COMBINATION OF POOLING
STRATEGIES, QEA & SEQ-QEA

↓

ARRIVE AT PAIRS OF INTERACTIVE PROTEINS UNIQUE TO A POPULATION

↓

SCREEN FOR INHIBITORS OF THESE UNIQUE INTERACTORS USING
A YEAST-BASED SELECTION OF INHIBITION OF INTERACTIONS

↓

ARRIVE AT LEAD DRUG COMPOUNDS THAT ARE SPECIFICALLY TARGETED
AGAINST PAIRS OF INTERACTING PROTEINS UNIQUE TO A PARTICULAR POPULATION

FIG. 1

D = DNA-BINDING DOMAIN

A = ACTIVATION DOMAIN

X = DNA-BINDING DOMAIN FUSION
      PROTEIN

Y = ACTIVATION DOMAIN FUSION
      PROTEIN

GROWTH ON MEDIA THAT
SELECTS FOR INTERACTING
PROTEINS
(e.g. -HIS or -URA MEDIUM)

MONITORING β-GALACTOSIDASE
ACTIVITY OF INTERACTING
PROTEINS

FIG. 2

M x N SCREEN

M & N ARE TWO POPULATIONS
OF PROTEINS SPECIFIC TO
A PARTICULAR STATE
(e.g., CANCER)

GROWTH ON -URA -HIS +3-AT MEDIUM (SELECTS FOR INTERACTORS)
AND SCREEN FOR β-GALACTOSIDASE ACTIVITY

CREATION OF INTERACTIVE POPULATION

TWO- OR THREE DIMENSIONAL POOLING STRATEGIES
TO FORM POOLS OF ALL INTERACTING PAIRS

YEAST COLONY PCR FOR M & N INSERTS ON POOLS

M-PCR   N-PCR

QEA & SEQ-QEA

CHARACTERIZE PAIRS OF INTERCTING PROTEINS IN ONE DISEASE-STATE

PERFORM ANOTHER M x N ANALYSIS FOR A DIFFERENT STATE (eg., NORMAL)

COMPARISON OF THE TWO M x N ANALYSES LEADS TO THE IDENTIFICATION
OF PAIRS OF INTERACTING PROTEINS SPECIFIC TO CERTAIN STATE (eg., CANCER)

FIG. 3

A.

B.

C.

FIG. 4

FIG. 5

N CHEMICALS x M INTERACTANTS

N x M SCREENS

NORMAL x NORMAL　　　　CANCER x CANCER　　　　NORMAL x CANCER

POOL ALL POSITIVE INTERACTANTS (COLONIES)

EACH ● REPRESENTS THE POOL OF STAGE-SPECIFIC INTERACTANTS

SCREEN FOR INHIBITION BY SMALL ORGANICS USING THE
INHIBITION-DEPENDENT GROWTH ASSAY

NORMAL x NORMAL　　　　CANCER x CANCER　　　　NORMAL x CANCER

AT THIS STEP EACH ● RECEIVES A DIFFERENT CHEMICAL

● INHIBITION DEPENDENT GROWTH

O NO GROWTH (NO INHIBITION)

AT THIS STEP THE INHIBITORS
USED CAN BE ENCODED SO AS TO
INCREASE CHEMICAL DIVERSITY
(COMBINATORIAL APPROACHES
CAN BE USED HERE)

ARRIVE AT STAGE-SPECIFIC (NORMAL/DISEASE) INTERACTING PAIRS OF
PROTEINS AND SMALL MOLECULAR INHIBITORS OF THESE INTERACTIONS

FIG. 6

ADC1-P   PEPTIDE   NLS   UAG ADC1-T

      Sfi I   Asc I

FIG. 7

FIG. 8

FIG. 9

FIG. 10A

5'- AGC ACT CTC CAG CCT CTC ACC GAA-3' (SEQ ID NO: 1)

250   3'- AG TCG CTT CTAG-6' (SEQ ID NO: 7)

5' ACC GAC GTC GAC TAT CCA TGA AGC-3' (SEQ ID NO: 42)

251   3'- GT ACT TCG TCGA-5'   (SEQ ID NO: 44)

FIG. 10B
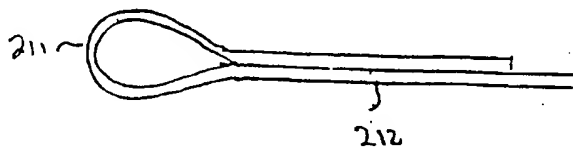
FIG. 10C
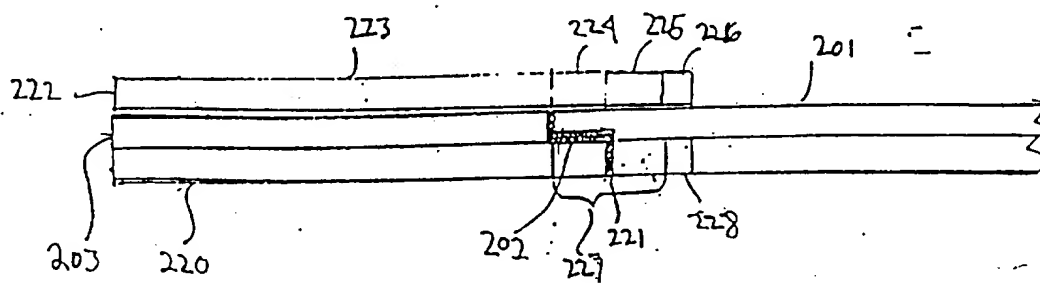
FIG. 10D
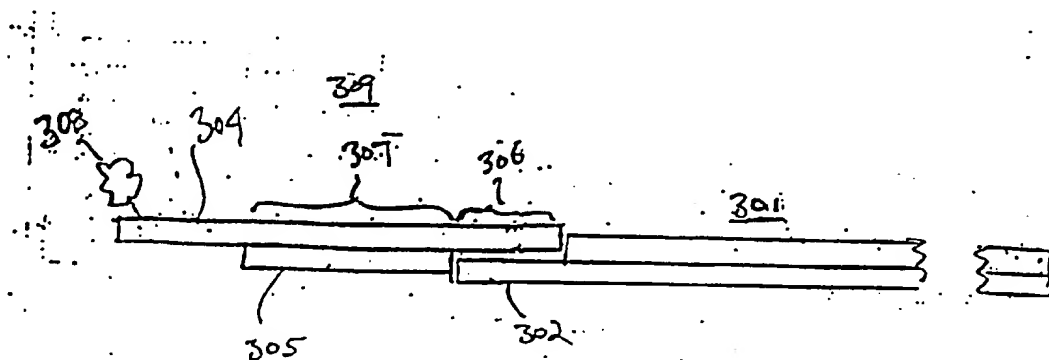
FIG. 11A

5'- AGCACTCTCCAG.CCTC.TCAC.CGAGCATG   (SEQ 10 NO.56)
      3'- AGTG GCTC

FIG. 11B

FIG. 12A

FIG. 12B

FIG. 12C

FIG. 13A

FIG. 13B

FIG. 14

FIG. 15A

FIG. 15B

FIG. 16

FIG. 17A

1401   1403
V1 = ($[10,14]_{RE1}$, $[62,66]_{RE1}$, $[610,614]_{RE1}$)

1402
V2 = ($[237,241]_{RE2}$, $[388,392]_{RE2}$)

VECTORS OF
CUTS

FIG. 17B

1404   1405   1406   1407
($[10,14]_{RE1}$, $[62,66]_{RE1}$, $[237,241]_{RE2}$,

        $[388,392]_{RE2}$, $[610,614]_{RE1}$)

MERGED AND
SORTED · VECTORS
OF CUTS

FIG. 17C

1408   1409
[RE1, RE1, *, 52]   [RE1, RE2, *, 75]   [RE2, RE2, *, 151]

                                             [RE1, RE2, *, 222]

FRAGMENTS

FIG. 17D

1410
[10,62]   [62,237]   [237,388]   [388,610]

FRAGMENT
SEQUENCE

FIG. 17E

A11
({ [RE1, RE1, N, 52], [RE2, RE2, Y, 151],
    [RE1, RE2, N, 175], [RE1, RE2, N, 222] })
                1412

SORTED FRAGMENT
VECTOR

FIG. 17F

DIGEST
TABLE

	[RE1, RE1]	[RE1, RE2]	[RE2, RE2]
:			
52	+A01		
:			A12
151			+A01
:			
175		+A01	
:			
222		+A01	
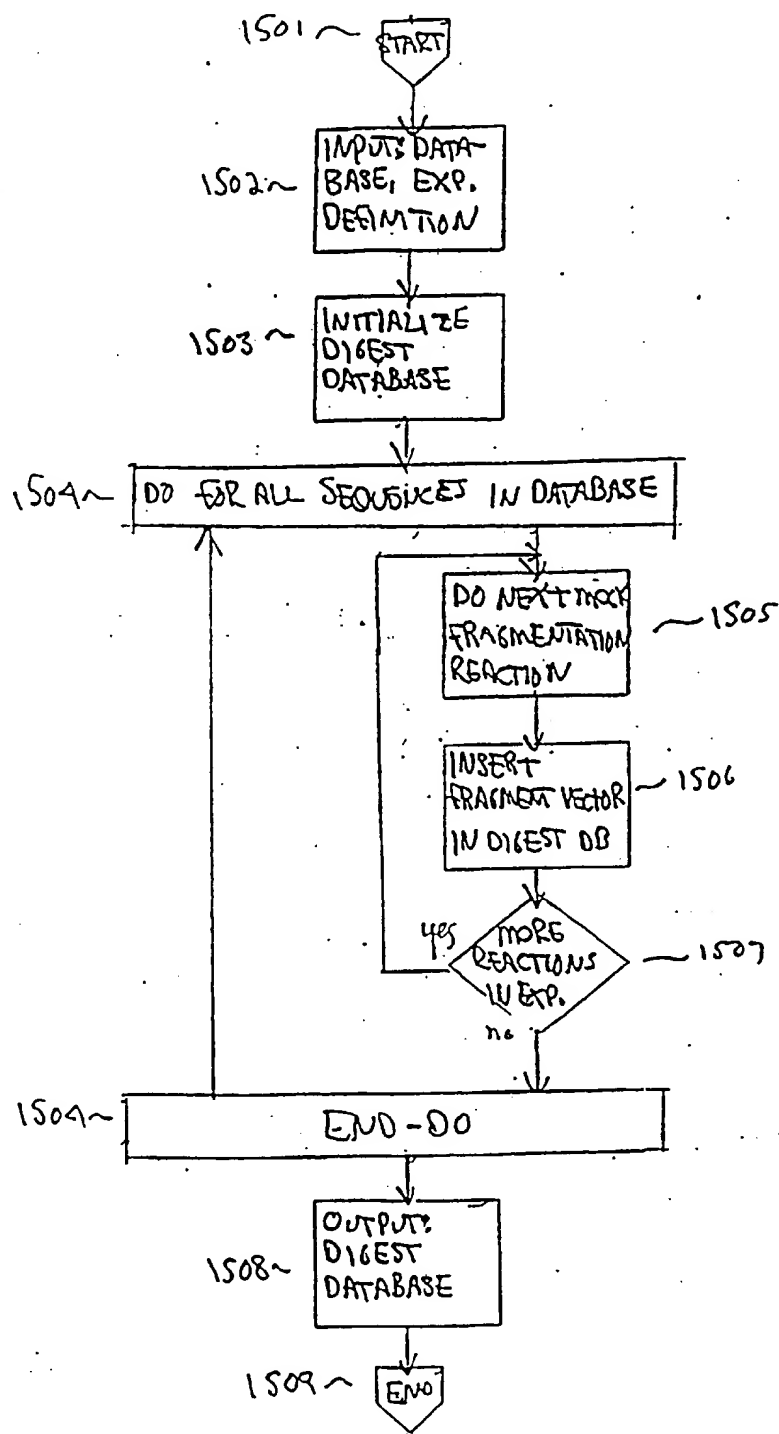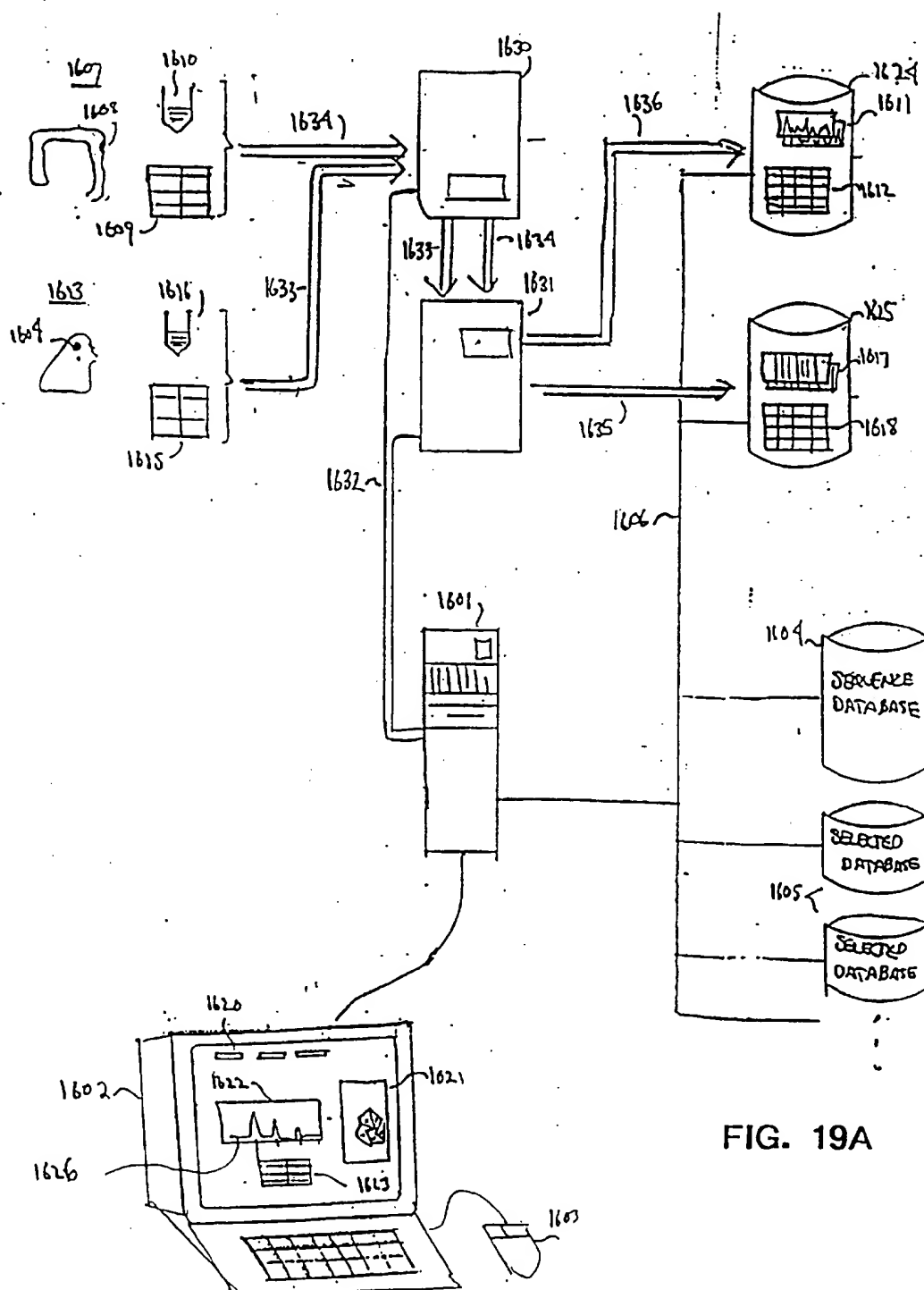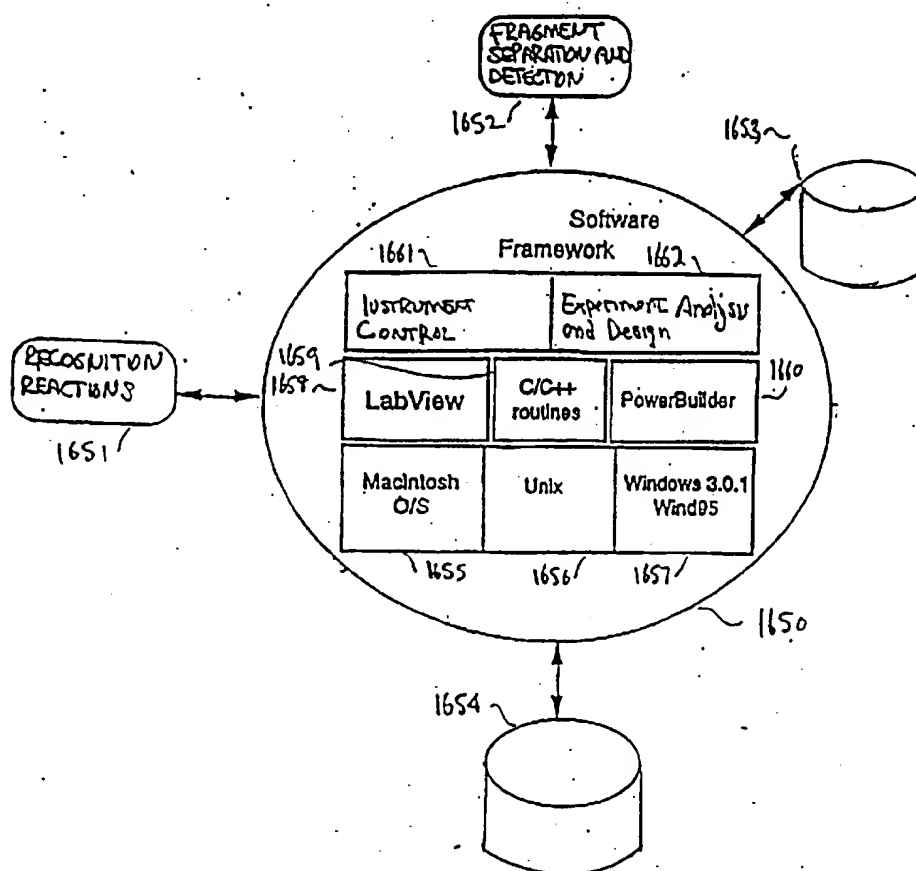
FIG. 18

FIG. 19A

FIG. 19B
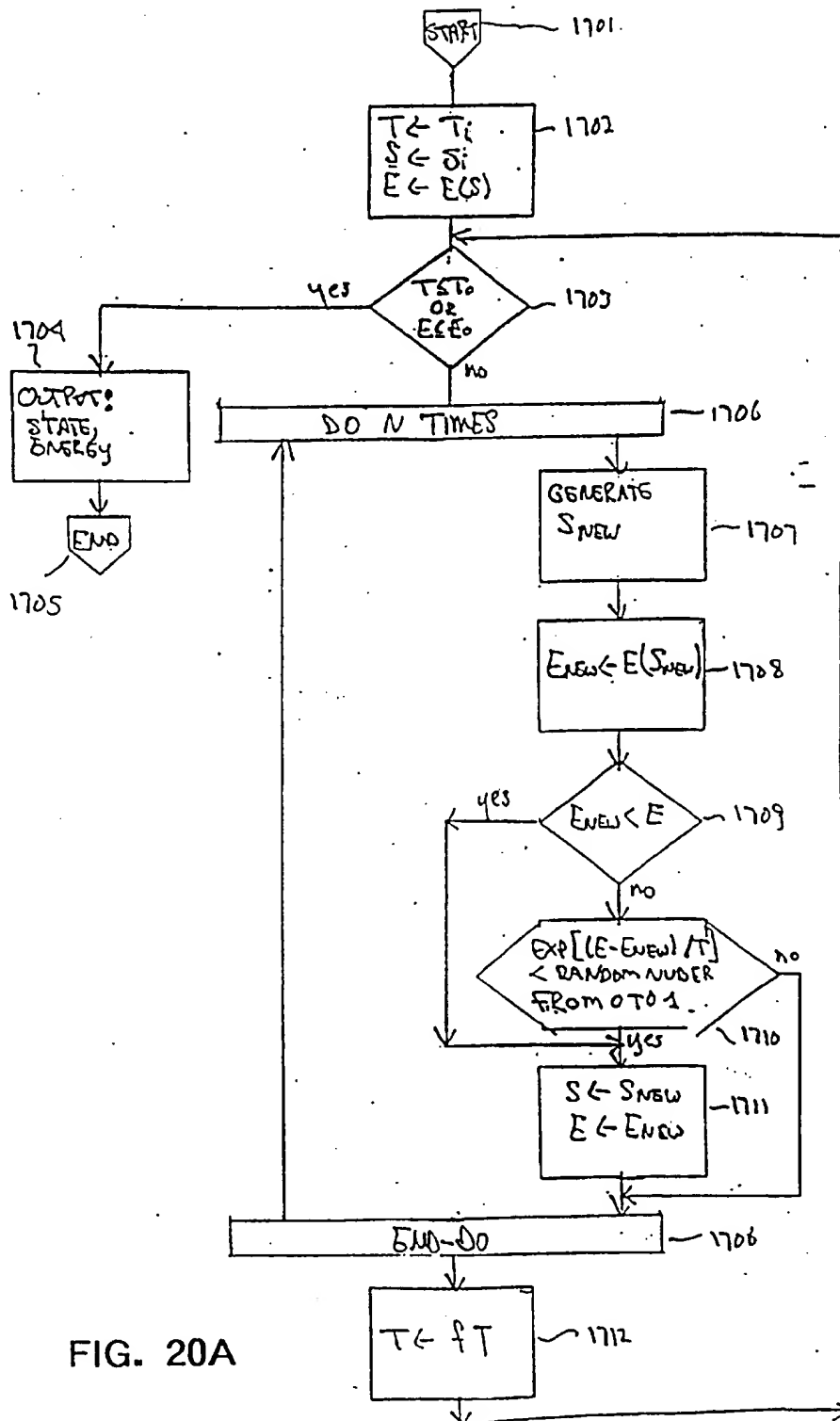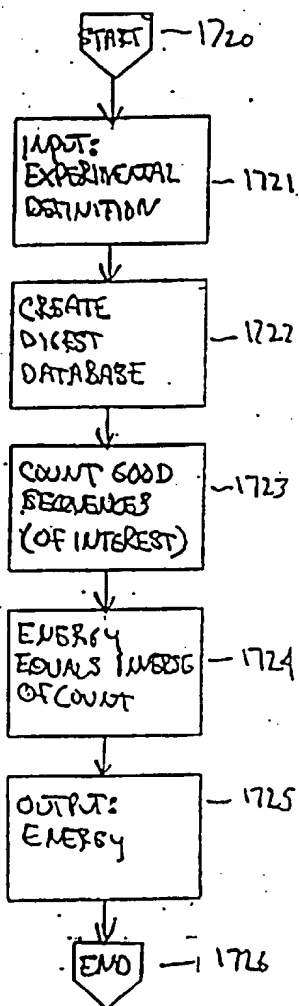
FIG. 20A

FIG. 20B
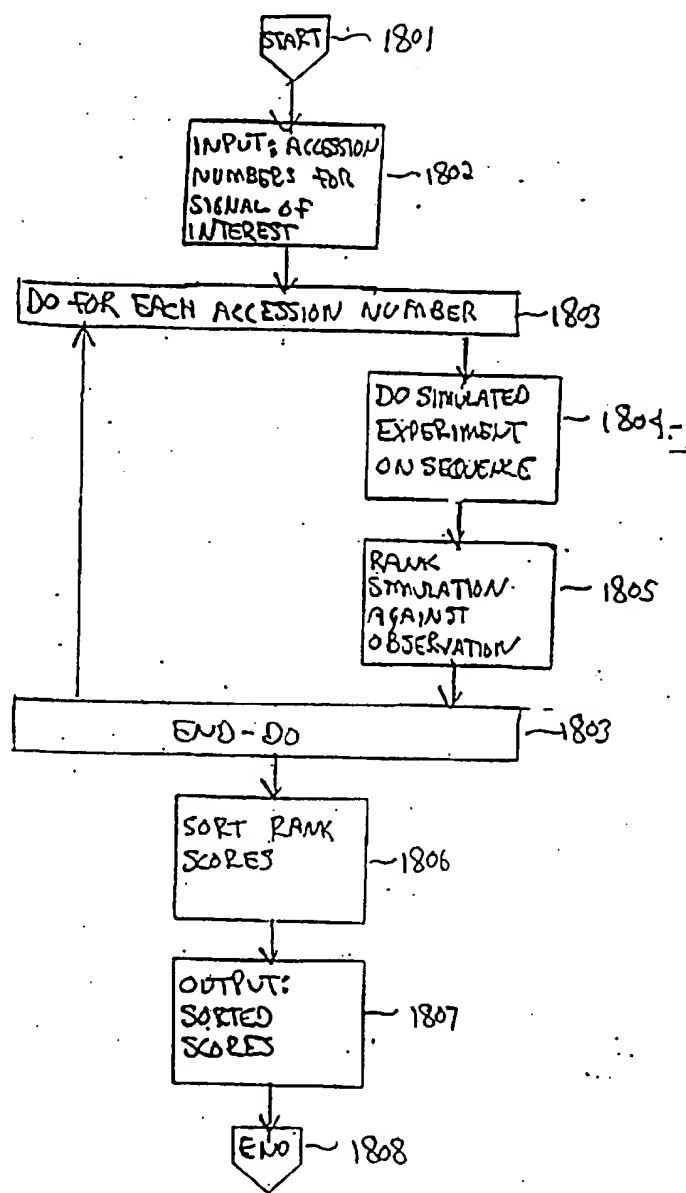
FIG. 21
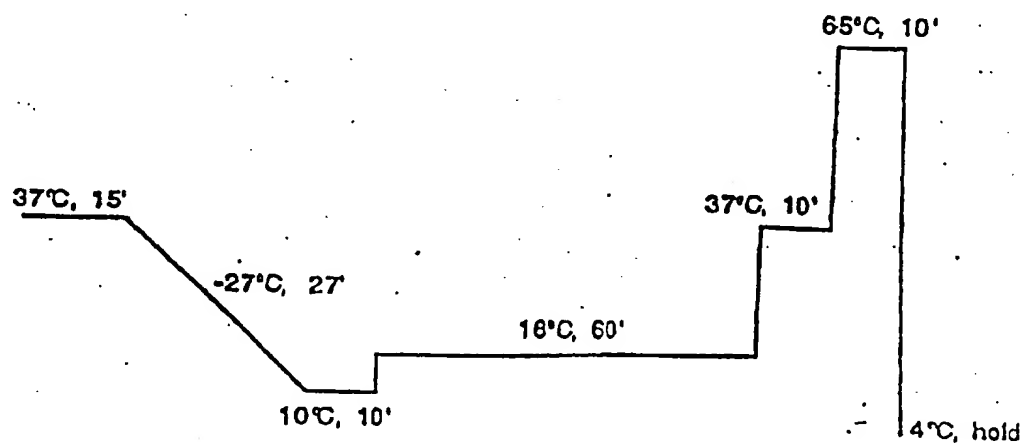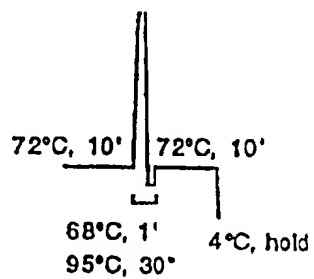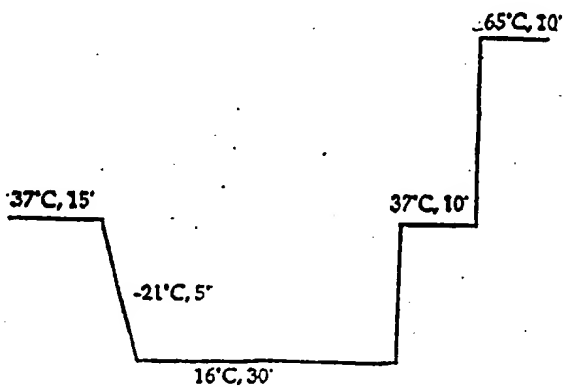
FIG. 22A

FIG. 22B

FIG. 22.C

FIG. 22D

FIG. 23A

FIG. 23B

FIG. 23C

FIG. 23D
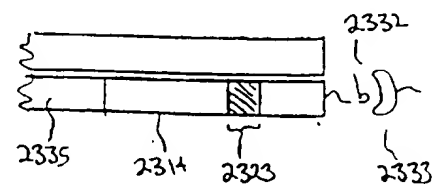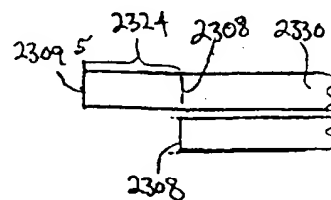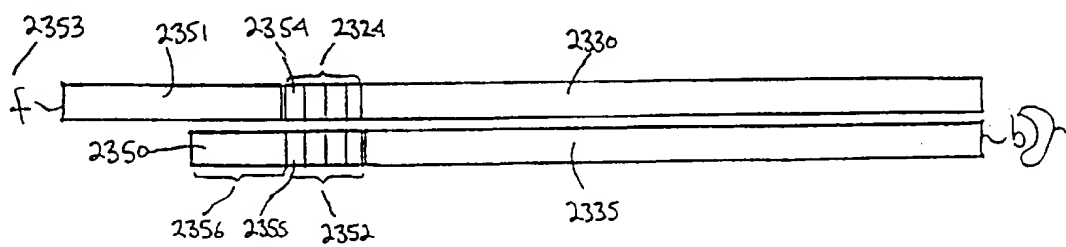
FIG. 23E

ISOLATION OF INTERACTING PROTEINS FROM AN M X N ANALYSIS

ISOLATE URA+, HIS+, lacZ+, CELLS: POSITIVE FOR PROTEIN-PROTEIN INTERACTION

eg., VEGF-VEGF, R4-FKBP-12, RAS-RAF etc.,

$\downarrow$

TRANSFER TO AND GROWTH IN NON-INDUCING MEDIA (LACTATE)

$\downarrow$

TRANSFER TO AND GROWTH IN INDUCING MEDIA (GLUCOSE) THAT
IS SELECTIVE FOR INTERACTION (-URA), ALONG WITH INHIBITOR

$\downarrow$

TRANSFER TO AND GROWTH IN 5-FOA MEDIA, ALONG WITH INHIBITOR

$\downarrow$

SELECTION OF 5-FOA^R CELLS: CELLS IN WHICH PROTEIN-PROTEIN INTERACTIONS HAVE
BEEN INHIBITED

eg., SELECTION OF R4-FKBP12 CELLS IN THE PRESENCE OF FK506 IN 5-FOA

FIG. 24

ISOLATION OF INTERACTING PROTEINS FROM AN M X N ANALYSIS

- ISOLATE URA+, HIS+, lacZ+, CELLS: POSITIVE FOR PROTEIN-PROTEIN INTERACTION

      eg., VEGF-VEGF, R4-FKBP-12, RAS-RAF etc.,

SELECTING INHIBITON OF NOVEL PROTEIN-PROTEIN INTERACTIONS BY CANDIDATE
INHIBITORS USING THE 5-FOA ASSAY

- POOL ALL INTERACTANTS FROM AN M X N ANALYSIS

      eg., R4-FKBP12 IN A POOL OF INTERACTANTS

- SCREEN AGAINST INHIBITORS USING THE 5-FOA ASSAY

      eg., FK506 AGAINST R4-FKBP12 IN A POOL OF INTERACTANTS

- SELECTION OF THOSE PROTEIN-PROTEIN INTERACTION EVENTS
WHERE INHIBITION OCCURRED

      eg., SELECTION OF R4-FKBP12 AS 5-FOA RESISTANT CELLS AMONG A POOL
OF INTERACTANTS WHEN EXPOSED TO FK506

ISOLATION AND PROTEIN-PROTEIN INTERACTIONS AND INHIBITORS OF THESE
INTERCATIONS

- CHARACTERIZATION OF THE GENES ENCODING THE INTERACTING PROTEINS
BY SEQUENCE ANALYSIS

- CONFIRMATION OF INHIBITION BY ENZYMATIC ASSAYS

      eg., β-GALACTOSIDASE ASSAYS

FIG. 25

INTERACTION METHOD STEPS   INFORMATION PROCESSING STEPS

CONSTRUCT FUSION LIBRARIES

2601

TRANSFORM YEAST STRAINS

2602

NEGATIVE SELECTION STEP

2603

MATING & ISOLATION OF
INTERACTANTS

2604

2614

2608   SEQUENCE FILES   2615
(QEA SIGNALS) &
CHARACTERIZE   MATING EXPERIMENT
FUSION PROTEINS   DESCRIPTION FILES   SEQUENCE
(PCR AND SEQUENCING)   DATABASES

2605

2609   GENE IDENTIFICATION   2616

2617   IDENTIFICATION   LOOKASIDE
DATABASE   DATABASES   2618

INTERACTION QUALITY
2610   CONTROL   2619

2611   2620
PLASMID DROP-OUT TEST   INTERACTION DATABASE
2606   &/OR MATRIX MATING TEST

2607   FURTHER STEPS   2612   DATABASE
BROWSING

2621

FIG. 26

PATHWAY
CONSTRUCTION

DOMAIN   2623
IDENTIFICATION

2622

FIG. 27

FG Study #1

Screen	Name	Bait Library	Prey Library	Status
30151	6xN	6 cloned Baits	Human Fetal Brain	Done
30152	18xN	18 cloned Baits	Human Fetal Brain	Done
30152a	18xN	18 cloned Baits	Human Fetal Brain	Done
30153	Mx22	22 cloned Baits	Human Fetal Brain	Done
30154	15xN	15 cloned baits	Human Fetal Brain	In Process
30154_b	15xN	15 cloned Baits	Human Fetal Brain	In Process
30155	Mx20	20 cloned baits	Human Fetal Brain	In Process
30156	25xN	25 cloned Baits	Human Fetal Brain	In Process

FIG. 28A

PathMaker™ Filter

Confirmation Filter: O passes plasmid dropout O passes matrix mating O passes both ⊙ with or without confirmation	Screen Filter: O forward screen only O reverse screen only O bi-directional screen ⊙ show all screens	Source Filter: O novel discovery O taken from the literature O from literature/confirmed in-house ⊙ display with either source
[List] as O proteins ⊙ interactions O proteins & interactions	Threshold of Independent Isolates: Minimum # of isolates: [1]	Submit to PathMaker™ with these filters: [Graph]

**FIG. 28B.**
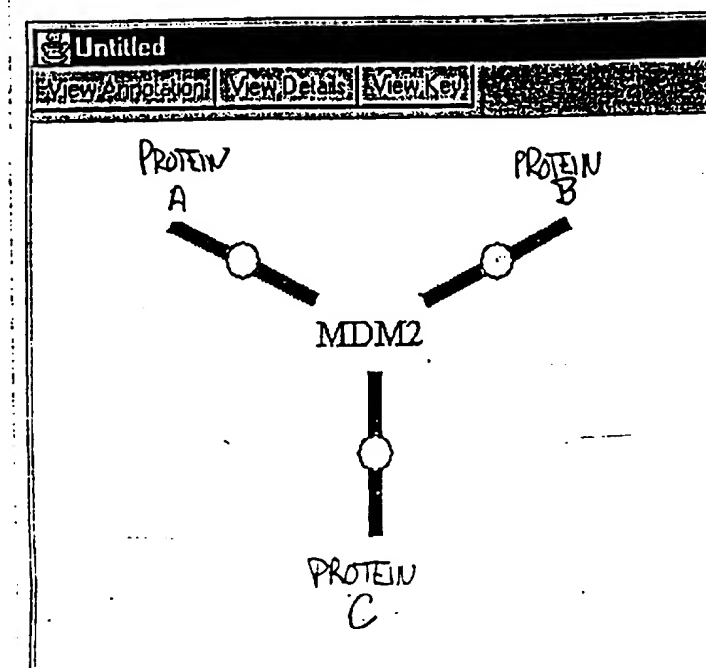
FIG. 29
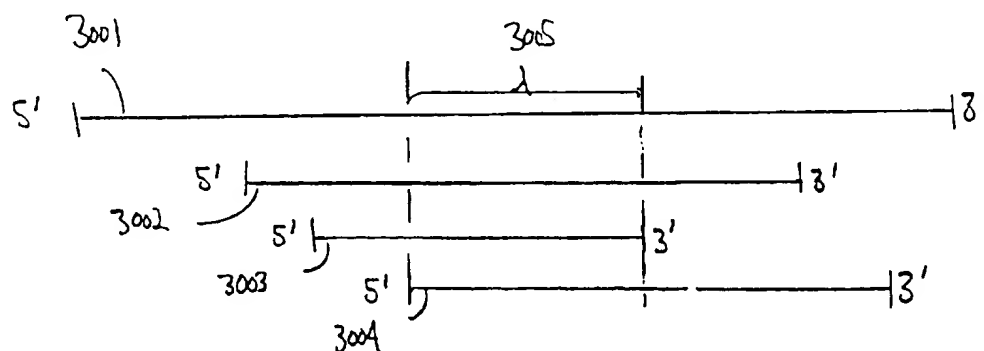
FIG. 30

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US97/10392

A. CLASSIFICATION OF SUBJECT MATTER
IPC(6) :Please See Extra Sheet.
US CL :Please See Extra Sheet.
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED
Minimum documentation searched (classification system followed by classification symbols)
U.S. : 435/69.1, 71.1, 254.2, 254.21, 320.1; 395/427, 616, 613, 601, 603, 606; 364/130, 413.01

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	BARTEL et al. A Protein Linkage MAp of Escherichia coli bacteriphage T7. Nature Genetics. January 1996. Vol. 12. Pages 72-77, see especially figure 1, page 73, page 76, last full paragraph, and page 77.	1-16, 21-47, 52-89, 91, 94, 95, 114-154, 180-190
X --- A	FINLEY, Jr. et al. Interaction Mating Reveals Binary and Ternary Connections between Drosophila Cell Cycle Regulators. Proc. Natl. Acad. Sci. USA. December 1994. Vol. 91, pages 12980-12984, see page 12984.	155-169 ----------- 170-179
X	US 5,525,490 A (ERICKSON et al.) 11 June 1996, see columns 3-7.	17-20, 48-51, 90, 92, 93, 96-113

☐ Further documents are listed in the continuation of Box C.   ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier document published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13 AUGUST 1997	1 5 SEP 1997

Name and mailing address of the ISA/US	Authorized officer
Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	Amy Atzel, Ph.D.
Facsimile No. (703) 305-3230	Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

	International application No. PCT/US97/10392

Box I   Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
   because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
   because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
   because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II   Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

   Please See Extra Sheet.

1. ☒ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest   ☐ The additional search fees were accompanied by the applicant's protest.

☐ No protest accompanied the payment of additional search fees.

Form PCT/ISA/210 (continuation of first sheet(1))(July 1992)✶

INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER:
IPC (6):

C12P 21/06, 21/04; C12N 1/14, 1/16, 1/18, 15/00, 15/09, 15/63, 15/70, 15/74; G06F 15/00

A. CLASSIFICATION OF SUBJECT MATTER:
US CL :

435/69.1, 71.1, 254.2, 254.21, 320.1; 395/427, 616, 613, 601, 603, 606; 364/130, 413.01

B. FIELDS SEARCHED
Electronic data bases consulted (Name of data base and where practicable terms used):

APS, CAPLUS, BIOSIS, EMBASE, INPADOC, MEDLINE, SCISEARCH, LIFESCI
search terms:
yest, two hybrid, GAL4, protein protein interaction, expression, DNA-binding, transcriptional activator, computer, sofware, database, sequence, nucleic acid, digital, Fields, Brent, pool

BOX II. OBSERVATIONS WHERE UNITY OF INVENTION WAS LACKING
This ISA found multiple inventions as follows:

This application contains the following inventions or groups of inventions which are not so linked as to form a single inventive concept under PCT Rule 13.1. In order for all inventions to be searched, the appropriate additional search fees must be paid.

Group I, claim(s) 1-16, 21-47, 52-85, 94, 95, 114-154, 180-190, drawn to methods of detecting protein-protein interactions in host cells.
Group II, claim(s) 86-89, drawn to cells and kits.
Group III, claim(s) 91, drawn to vectors.
Group IV, claim(s) 17-20, 48-51, 90, 92, 93, 96-113, drawn to methods of detecting inhibitors.
Group V, claim(s) 155-179, drawn to computer-implemented methods, computer systems, memories and software.

The inventions listed as Groups I-V do not relate to a single inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons: Groups I & IV do not share the same or a corresponding special technical feature with the computer system of Group V or the vector of Group III or the cells and kits of Group II; and the methods of Group I does not share a special technical feature with Group IV as it lacks the presence of inhibitors of Group IV.